Cue - H05438 - 4 - P5027

THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE

9

EDITOR
Paul W MacAvoy, *Massachusetts Institute of Technology*

ASSOCIATE EDITORS:
Dale W. Jorgenson, *Harvard University*
Harvey J. McMains, *American Telephone and Telegraph Company*
Oliver E. Williamson, *University of Pennsylvania*

ASSISTANT TO THE EDITOR:
Leigh E. Tripoli, *American Telephone and Telegraph Company*

SPRING '74
VOL. 5 NO. 1

cpt 1974

CENTRAL LIBRARY
UNIVERSITY

TABLE OF CONTENTS

ARTICLES

3 Regulation of natural monopolies and the fair rate of return
H. E. LELAND

16 A theory of interdependent demand for a communications service
J. ROHLFS

38 Rate of return regulation and efficiency in production: an empirical test of the Averch-Johnson thesis
R. M. SPANN

53 Regulation and efficiency in the electric utility industry
L. COURVILLE

75 Reversals in peak and offpeak prices
E. E. BAILEY and L. J. WHITE

93 The effect of regulation changes on insider trading
J. F. JAFFE

SYMPOSIUM ON THE OPTIMALITY OF COMPETITIVE CAPITAL MARKETS

125 Production theory and the stock market
H. E. LELAND

145 The optimality of a competitive stock market
R. C. MERTON and
M. G. SUBRAHMANYAM

171 On the theory of the firm in an economy with incomplete markets
S. EKERN and R. WILSON

181 A note on unanimity of stockholders' preferences among alternative production plans: a reformulation of the Ekern-Wilson model
R. RADNER

SHORT ARTICLES, COMMENTS AND REVIEWS

187 Space-time pricing for public utilities
J. CRAVEN

196 Corporate diversification and market structure
C. H. BERRY

205 The CAB's domestic passenger fare investigation
G. W. DOUGLAS and J. C. MILLER, III

223 Peak-load pricing: some further remarks
O. E. WILLIAMSON

229 Depreciation in a simple regulatory model: comment
P. B. LINHART

233 Depreciation in a simple regulatory model: reply
B. L. JAFFEE

235 Cable, cities, and copyrights
B. M. MITCHELL in association with
R. H. SMILEY

264 A reexamination of the prophecy of doom for cable television
R. W. CRANDALL and L. L. FRAY

290 Cowling's *Market Structure and Corporate Behavior*
L. WEISS

294 Posner's *Economic Analysis of Law*
P. A. DIAMOND

301 Noll, Peck, and McGowan's *Economic Aspects of Television Regulation*
S. M. BESEN and B. M. MITCHELL

320 Approving nuclear power plants: scientific decisionmaking or administrative charade?
P. L. JOSKOW

P-5027

Regulation of natural monopolies and the fair rate of return

Hayne E. Leland

Assistant Professor of Economics
Stanford University

Defining "fair return" when profits are random has been a central problem in the theory of regulation. In this paper, we develop a simple but general concept of fair return based on an equilibrium model of production under uncertainty. We propose regulatory behavior which will induce firms to make efficient input choices, while at the same time guaranteeing fair returns.

1. Introduction

■ Regulation is a common phenomenon in "natural monopolies," those industries with decreasing average costs. Optimal resource use in such industries dictates that output be provided by a single firm, but this easily leads to inefficient monopolistic practices by firms in the absence of regulation. Effective regulation, however, requires that meaningful regulatory objectives be specified and rules devised for their implementation. While economic theory provides appropriate objectives and behavioral rules when there is no uncertainty, the presence of uncertainty is central to the knottiest problems of regulation. And until recently, economic theory had little to say about optimal production in the presence of uncertainty.

A particularly crucial problem facing regulatory agencies is an appropriate definition of a "fair" rate of return. The legal guideline for regulation, based on the *Hope* decision, requires that regulatory policies provide a return to the equity holder "commensurate with returns on investments with corresponding risks."[1] Under certainty, this guideline is well-defined: equity holders should receive *the* riskless rate of return which prevails throughout the economy. Under uncertainty, the guideline is much less clear. Often corresponding risks are not present. And if they are present, they usually are associated with other regulated firms, leaving a simultaneity problem. In short, economic theory has not provided a definitive concept of "fair" return under conditions of risk, and this lack of clarity has led to costly hearings and litigation "with

---

Hayne E. Leland received the A.B. from Harvard (1963), the M.Sc. (Econ.) from the London School of Economics (1965), and the Ph.D. from Harvard (1968). His current research deals with the economics of uncertainty, the extraction of natural resources, and the behavior of the regulated company.

The research described in this paper was undertaken from grants from the National Science Foundation and from the Ford Foundation. The author thanks Alvin K. Klevorick and Oliver E. Williamson for guidance along the labyrinthine paths of regulatory theory and practice. Wrong turns, of course, remain the author's responsibility.

[1] *F.P.C. v. Hope Natural Gas Company,* 320 U.S. 591 (1944).

final results which are often inconsistent with the goal of optimal resource allocation in the economy."[2]

In this paper, we build on the production/stock market equilibrium model developed in our paper which appears in the Symposium on the Optimality of Competitive Capital Markets in this issue.[3] Our purpose is to construct a concept of fair return, and to suggest regulatory behavior consistent with the achievement of optimal resource use. Unique to our theory of fair return is the fact that, while profits are random variables, we do not need to measure subjective probabilities in order to set policies which generate fair returns. But stock market values, which reflect investors' expectations and attitudes towards risk, are essential to our theory.

The spirit of our approach is similar to that of Myers and of Gelhaus and Wilson,[4] who suggested that stock market values were important in assessing fair returns. But the theoretical underpinnings of those papers were not grounded on an equilibrium model of both production and the stock market, and Myers' theory of fair return required (1) estimates of expected returns to investment; and (2) estimation of "beta coefficients" as a possible guide to assessing the cost of capital. Our theory avoids both these difficulties by building on a model which considers equilibrium simultaneously in production and in financial markets.

2. Production and the stock market

■ In our Symposium paper[5] we develop a model of production and financial market equilibrium with the following properties:

(1) Investors are typified by heterogeneous initial wealths, utility functions, and expectations;
(2) Firms are typified by profit functions depending on the firms' decisions and the state of nature;
(3) Equilibrium is typified by investors choosing optimal portfolios, conditional on market clearing prices of stocks and on firms' production decisions; and by the production decisions of firms being "in the stockholders' interests," conditional on the market clearing prices and optimal portfolios of investors.

Key to our theory is the proof that stockholders of each firm unanimously recommend production decisions, when the stochastic formulation satisfies a weak restriction.[6] For firms whose returns are nonstochastic, stockholders will choose profit maximizing policies. For firms with stochastic returns, profit is not well defined, but equilibrium conditions will hold, given the policies chosen by stockholders. These conditions are exploited in Section 5.

---

[2] See Gelhaus and Wilson [3], p. 287.

[3] In [5].

[4] In [6] and [3], respectively.

[5] In [5].

[6] For unanimity, we require profit take the form $\Pi(d, \theta) = f(d) + g(d)h(\theta)$, where d is a vector of decisions. If profit does not take this form, stockholders would wish the firm to split up into firms whose profit does take this form. See [5], note 33.

We develop below a simple model of a monopolistic firm with uncertain returns. Some notation is inescapable. Let

q = Output.
$p(q)$ = Selling price.
K^0 = The number of units of capital owned by the firm. Capital units are presumed tradable at a market price normalized to one. Therefore, K^0 also equals the cost of the firm's assets.[7]
K^b = The amount of additional capital hired by the firm.
r = The (riskless) rental rate per unit of capital, and therefore equals the riskless rate of return.
$K = K^0 + K^b$ = the total amount of capital used by the firm.
L = The number of units of variable input hired by the firm.
$w(\theta)$ = The price per unit of the variable input, which depends on the state of nature θ. If there is no uncertainty, then $w(\theta) = \overline{w}$ for all θ.
$F(K, L)$ = The production function of the firm, relating output to input levels of K and L.
V = Total value of the firm's shares on the stock market.

For simplicity, we assume the firm has no debt.[8] Profit will be given by

$$\Pi = p(q)q - rK^b - w(\theta)L \qquad (1)$$

when there is uncertainty about $w(\theta)$, and by

$$\Pi = p(q)q - rK^b - \overline{w}L \qquad (2)$$

when there is no uncertainty. Of course, (2) is a special case of (1), and results we derive for the case with uncertainty will also hold for the certainty situation.

Functional form (1) introduces uncertainty in a very simple way: the cost of input L is considered random.[9] Nonetheless, other types of uncertainty are consistent with our conclusions, including output uncertainty, given price.[10]

---

[7] K^0 is the "true economic value" of the firm's assets, which may or may not be closely approximated by the book value of the firm. See our comments in Section 6 below. For what follows, we assume K^0 is fixed. See note 8, however.

[8] If firms issue debt, interest on it should be included in firm's cost functions, and K^0 would increase by the amount of debt D. In the analysis which follows, the presence of debt D would require that V (the stock market value of the firm) be replaced by $V + D$, the sum of current equity and debt values. It can easily be shown that the firm would be indifferent between renting capital K^b or borrowing an amount D at the same cost r and increasing K^0 by D. That is, any increase in K^0 through a change in D will be exactly offset by a (negative) change in K^b. Given risk-free debt, the Modigliani-Miller theorem further implies that a change in K^0 resulting from equity financing would have precisely the effect of a change in K^0 resulting from a change in D. Therefore, our analysis with K^0 fixed and K^b variable is, in fact, perfectly general. If there is a possibility of bankruptcy, the analysis becomes more complicated.

[9] While the form (1) no doubt is an oversimplification, fuel and labor costs seem to be an important source of uncertainty to public utilities.

[10] See note 6.

When a firm's profit function is independent of θ, indicating a riskless profit, our analysis in the Symposium paper[11] indicates that equilibrium requires firms to maximize profit and have total stock market value,

$$V = \Pi/r, \tag{3}$$

where Π is maximal profit and r is the riskless return. Equation (3) follows immediately from arbitrage considerations: if V does *not* satisfy (3), then a sure profit can be made by buying (or selling) the riskless asset (whose price is normalized to one) and selling (or buying) stock in the firm with the riskless profit Π. When profits involve risk, (3) is not a meaningful relation, and is replaced by different equilibrium requirements.[12]

3. Fair return

■ Legal precedent requires that regulatory policies assure the investor a rate of return which provides commensurate returns and is (just) "sufficient to attract capital to the firm."[13] Under certainty, this would be straightforward. Defining the rate of return as Π/K^0, a fair rate of return would be precisely the riskless interest rate which prevails throughout the rest of the system—namely, the rate r. Any higher return would be more than required to keep capital in its current use, whereas any less would lead to its withdrawal.

In the context of equilibrium in the stock market, the achievement of a "fair" rate of return has an interesting implication:

Proposition 1: For any riskless firm, a fair rate of return r implies the stock market value of the firm, V, is exactly equal to the value of its assets, K^0. If the firm earns an "excess" rate of return ($\Pi/K^0 > r$), the stock market value of the firm will exceed the value of its assets.

Proof. A "fair" rate of return implies $\Pi/K^0 = r$. From (3), $V = \Pi/r$. Together, these imply $V = K^0$. An excess return implies $\Pi/K^0 > r$, which with (3) implies $V > K^0$. Q.E.D.

Under uncertainty, profits depend on the state of nature, and therefore the ratio Π/K^0 is meaningless. If the expectations of all investors were identical and known to regulators, then $E(\Pi)$ would be meaningful. But it would still be necessary to define an r^* such that $E(\Pi)/K^0 = r^*$ would be the fair return to the firm, given the extent of uncertainty. Myers[14] works from this point, making strong assumptions about investors' preferences and expectations. Fortunately, there is a simpler way to determine a fair return.

For firms facing either certain or uncertain profits, there may be no short-run relation between the market value of a firm's stock, V, and the value of the assets it owns, K^0. But, if the firm were in a competitive environment with freedom of entry and exit, we would expect these values to converge. If $V > K^0$, entrepreneurs could make a sure profit by buying K^0 worth of assets, then selling stock for V. If $V < K^0$, the entrepreneur could make a sure profit by buying the firm for V and selling its assets for K^0.[15] Entry and/or exit will eventually drive the patterns of profits across states of nature to the point where $V = K^0$. In the case of riskless firms, Proposition I demonstrates that this implies "zero" long-run profits, where costs include a fair return r to capital.

---

[11] In [5].

[12] See [5], note 32.

[13] *F.P.C. v. Hope Natural Gas Company*, 320 U.S. 591 (1944).

[14] In [6].

At the long-run equilibrium with $V = K^0$, the pattern of profits is just sufficient to attract capital to its current use. Our argument then leads to a definition of "fair return."

Definition: A "fair return" to capital is a pattern of profits across states of nature just sufficient to attract capital to its present use, which is equivalent to the stock market value of the firm, V, equalling the value of the firm's assets, K^0.

We might note two things:

(1) Our definition coincides with the usual definition of fair return under certainty, by Proposition I;

(2) Our definition is operable in the presence of uncertainty without requiring homogeneity or knowledge of subjective expectations. Thus, it avoids the pitfalls inherent in Myers' definition.[16]

We postpone a discussion of the implications (and difficulties) inherent in our definition until Sections 5 and 6.

4. Regulation under certainty

■ Before a meaningful discussion of appropriate regulatory behavior can be undertaken, we need a well-defined objective. Conventional wisdom, based on economic theory, has generally held that regulation should seek to achieve a Pareto optimal or "efficient" use of resources in the industry in question. Efficient use of resources under certainty can be shown to require (1) a technically or allocationally efficient use of inputs given the level of output, which requires cost minimization in the presence of perfect input markets; and (2) an efficient level of output, which requires that price equal marginal cost of production when inputs are optimally used.[17]

Unregulated monopolies which maximize profits typically will achieve (1), but not (2): monopolists produce too little from the point of view of social welfare, and charge too much. Figure 1 indicates the profit maximizing output q_M of the monopolist, where the cost curve includes a fair return r to owned assets K^0. The presence of positive (i.e., excess) profits at $q = q_M$ implies by Proposition I that the market value of the firm's stock, V, exceeds the value of its assets, K^0.

---

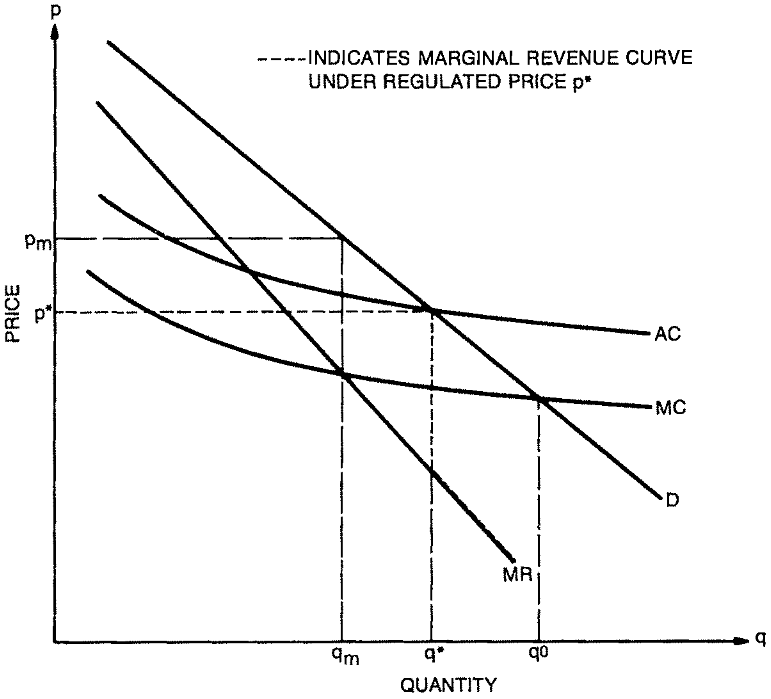
[15] (1) and Proposition I imply that if the firm can rent all its owned capital K^0, at rate r, it will be guaranteed a sure profit rK^0, and therefore a minimum value $V = K^0$. But presumably not all capital can be rented out —we assume $K_{\min}^b > -K^0$, implying the possibility of $V < K^0$. For simplicity, we assume the optimal K^b will occur in the interior of the feasible set.

[16] In [6], p. 80.

[17] See Scitovsky [7] or other texts for these results.

FIGURE 1

REVENUE AND COST CURVES UNDER CERTAINTY



But the achievement of the Pareto optimal output, q^0, involves a difficulty. Natural monopolies have declining average cost curves, which imply that marginal cost lies below average cost. Hence at q^0 the firm is not earning a fair return r on its investment K^0. Forcing the firm to produce at q^0 without subsidization would be in clear conflict with the *Hope* precedent, which requires that investors receive a fair rate of return.

The accepted "second best" goal for regulatory agencies, which can be justified on the basis of maximal net benefits,[18] seems to be the following:

> Regulatory Objective: the regulatory authority should seek the maximal output (lowest price) consistent with (1) a technically efficient use of inputs, and (2) a fair return to investment. (4)

Since technical efficiency is a well-defined concept under cost uncertainty, and since we have defined a notion of fair return under uncertainty, the regulatory objective (4) is precisely defined under uncertainty as well as under certainty. But under certainty, we have the further knowledge that requirement (1) implies inputs must minimize costs of production, and that requirement (2) implies that $\Pi/K^0 = r$.

We now demonstrate the (well-known) assertion that regulatory objectives can be achieved under certainty, if the regulatory agency sets the lowest price ceiling which is consistent with $V = K^0$. In Section 5, we show this same behavior will achieve the regulatory objectives under uncertainty.

---

[18] Net benefits are measured as the sum of producer's and consumers' surplus. See R. Willig, "Consumers' Surplus: A Rigorous Cookbook," IMSSS Technical Report No. 98, Stanford University, April 1973, for a rigorous consideration of this criterion.

Assume the regulatory agency sets a maximal selling price $\overline{p}$. Demand at that price will be given by $q(\overline{p})$, and it is easily shown that, in the presence of decreasing average costs, the natural monopoly will maximize profits by fully meeting this demand. For any $\overline{p}$, total revenues $\overline{p}q(\overline{p})$ will be fixed. Acting in stockholders' interests, the firm will maximize profits by minimizing the cost of producing $q(\overline{p})$. But cost minimization is well-known to imply, under certainty, that

$$F_K/F_L = r/\overline{w}, \qquad (5)$$

where F_K is the marginal product of capital $\partial F(K, L)/\partial K$, and F_L is the marginal product of labor. Since input decisions by other profit maximizing firms also will satisfy condition (5), the ratios of marginal products of the inputs will be identical for all firms and technical or allocative efficiency will be achieved. A fair rate of return will be guaranteed by the selection of $\overline{p} = p^*$ in Figure 1. At $q(p^*)$, average cost (including a fair rate of return to K^0) equals p^*. Excess profits are zero, and by Proposition I, market value $V = K^0$, which by definition implied a fair rate of return to investment K^0.

Quite contrary to observed regulatory behavior,[19] some authors have suggested that regulatory agencies fix a rate of return $s \geqq r$, but do not set price ceilings.[20] We can readily demonstrate that, under certainty, this approach is poor prescription as well as poor description. If the maximal return s is set equal to r, the firm will not earn an excess return, but there is no assurance of technical efficiency (or maximal output), since profit maximization is consistent with an infinite number of combinations of K and L.[21] If, on the other hand, s is set above r, the firm will earn an excess rate of return, it will choose an inefficient combination of inputs, and it will produce at less than the optimal output $q(p^*)$.[22]

5. Regulation under uncertainty

■ In the previous section, we showed that, under certainty, the regulatory objective (4) could be achieved by

> Regulatory Behavior: the regulatory agency should set as a price ceiling the least price of output consistent with the stock market value of the firm equalling the value of its assets $(V = K^0)$.[23] (6)

The proof under certainty was simplified by the fact that technical efficiency is equivalent to cost minimization, and the fair rate of return condition implies $\Pi/K^0 = r$. Neither simplification follows under uncertainty, since costs and profits are random variables. Nonetheless, we can still prove the following:

---

[19] See Klevorick [4].

[20] See [1], [2], and [8], among others. Note, however, that if p^* depends on the firm's decisions, and this feedback is perceived by the firm, the A-J model [1] may be a more realistic description.

[21] See Zajac [8].

[22] See Baumol and Klevorick [2].

[23] If the firm has debt D outstanding, then the price should be chosen so $V + D = K^0$, or $V = K^0 - D$; see note 8.

Proposition II: Regulatory behavior (6) achieves the objective (4) when firms act in the interest of stockholders, and (random) profits assume the form (1).

Proof. Let j index firms. We note that for any increasing, strictly quasi-concave production function $F^j(K^j, L^j)$, we can derive a function

$$L^j = L^j(q^j, K^j),$$

which gives the amount of labor required to produce q^j with capital stock K^j. It is easily shown that

$$\frac{\partial L^j(q^j, K^j)}{\partial K^j} \equiv L_K{}^j(q^j, K^j) = -F_K{}^j/F_L{}^j. \qquad (7)$$

Assume that regulation (or in the nonregulated sector, stockholders' interests) has dictated some price $\bar{p}^j$ and output $q^j(\bar{p}^j)$. When r and $w(\theta)$ are the same for all firms, and firms' returns take the form (1), profits will be given by

$$\Pi^j = \bar{p}^j q^j(\bar{p}^j) - rK^{jb} - w(\theta)L^j[q^j(\bar{p}^j), K^j]. \qquad (8)$$

Different choices of K^{jb} will produce different patterns of Π^j across states of nature. Following the lines of our Symposium paper,[24] we assume the choice of K^{jb} will be "in the stockholders' interests." In equilibrium, this was shown to imply

$$\frac{\partial E_i[U_i(R_i, \theta)]}{\partial K^{jb}} = E_i\{U'_i(R_i, \theta)[-r - w(\theta)L_K{}^j]\} = 0, \quad (9)$$

for all i and j, where i indexes investors, and R_i is the (random) return to the investor's optimal portfolio.

We can rewrite (9) as

$$L_K{}^j = \frac{-E_i[U'_i(R_i, \theta)]r}{E_i[U'_i(R_i, \theta)w(\theta)]}. \qquad (10)$$

The right-hand side of (10) is independent of j (and, by the Unanimity Theorem in our Symposium paper,[25] is independent of i), and therefore (10) and (7) imply

$$-L_K{}^j = F_K{}^j/F_L{}^j = F_K{}^l/F_L{}^l = -L_K{}^l \text{ for all } j, l. \qquad (11)$$

But (11) is well-known to be the requirement for *technical efficiency*.[26]

If p is set very high, say at p^h, assume monopoly profits will exist, i.e. $\widehat{V}^h > K^0$ at p^h. If p is very low, say at p^l, the firm will have a low value $V^l < K^0$.[27] Assuming that V is continuous in p implies that there exists a $p^* \epsilon (p^h, p^l)$ such that $V = K^0$ at p^*, i.e. p^* generates a fair return to capital. The lowest p^* such that $V = K^0$

---

[24] See [5].

[25] See [5].

[26] Note if $w(\theta) = w$ for all θ, the certainty case, then (11) reduces to the familiar

$$F_K/F_L = r/w. \qquad \text{(i)}$$

[27] See note 13.

will be the optimum price ceiling, since $q(p^*)$ will be the largest output consistent with technical efficiency and a fair rate of return. Q.E.D.

Certain properties of input decisions at the regulatory optimum are also of interest.

Proposition III: At the optimal regulatory price p^*, the following equilibrium relationship holds:

$$F_K/F_L = rL/[p^*q(p^*) - rK]. \tag{12}$$

Proof. From the portfolio equilibrium conditions in equation (8) of our Symposium paper,[28] $E_t[U'_t(R_t, \theta)(\Pi - rV)] = 0$, or using (1) and $V = K^0$,

$$E_t\{U'_t(R_t, \theta)[p^*q(p^*) - rK - w(\theta)L]\} = 0.$$

Solving for $E_t[U'_t(R_t, \theta)w(\theta)]$ and substituting into (9) gives

$$-L_K = \frac{rL}{p^*q(p^*) - rK},$$

which with (7) implies (12). Q.E.D.

Note that if the firm faces a certain price w for the variable input L, $p^*q(p^*) - rK^b - wL = rK^0$, or $p^*q(p^*) - rK = wL$. Substituting this into (12) gives $F_K/F_L = r/w$.

When expectations are homogeneous and firms' profits in equilibrium are nonnegatively correlated (as would be expected when $w(\theta)$ is the sole source of uncertainty), we can prove stronger results.

Proposition IV: Let firm A face a random price of input L, $w(\theta)$, with $E[w(\theta)] = w$, and let an identical firm B face a known price of input L, equal to the w above. Then, under optimal regulation:

(1) The ratio F_K/F_L will be higher for firm B than firm A.
(2) The price $p_A{}^*$ set by the regulatory authority will exceed $p_B{}^*$
(3) Firm A will produce a lower output under regulation.

Proof. (1) For the riskless firm B, a fair return implies

$$\Pi_B = p_B{}^*q(p_B{}^*) - rK_B{}^b - wL_B = rK^0,$$

where $K_B{}^b$ and L_B are the cost minimizing inputs given $p = p^*{}_B$. Cost minimization implies

$$\frac{F_K}{F_L} = \frac{r}{w}, \text{ for firm } B. \tag{13}$$

In equilibrium, it is easily seen that risk aversion on the part of investors will require a higher expected profit from firm A to generate the same market value (which under optimal regulation must equal K^0 for both firms).[29] That is,

---

[28] See [5]

[29] By examining first-order conditions associated every individual's portfolio selection problem, we can easily show that demand for an asset will be positive only if $E(\Pi)/V > r$, when the asset is positively correlated

$$E(\Pi^A) = p_A{}^*q(q_A{}^*) - rK_A{}^b - wL_A > rK^0,$$

or

$$p_A{}^*q(p_A{}^*) - rK_A - wL_A > 0,$$

or

$$p_A{}^*q(p_A{}^*) - rK_A > wL_A.$$

Introducing the above into (13) gives

$$\frac{F_K}{F_L} < \frac{r}{w} \text{ for firm } A, \tag{14}$$

which with (13) proves Proposition IV(1).

IV(2): follows immediately by noticing that for any $p \leqq p_B{}^*$, $E(\Pi^A) \leqq \Pi^B \leqq rK^0$ for any combination of inputs which will meet demand. Therefore $E(\Pi^A) > rK^0$ implies $p_A{}^* > p_B{}^*$.

IV(3): follows immediately from the presumption that the demand curve is downward sloping. Q.E.D.

A consequence of Proposition IV(1) is that firms which face uncertainty will not minimize expected costs, which requires $F_K/F_L = r/w$. This consequence is hardly surprising: when uncertainty affects the price of one factor and not the other, risk aversion leads to the substitution of the less risky factor (K for L in the assumptions of the model), despite some small increase in expected costs. It follows that the risky firm will use (for a given output) a higher K/L ratio than the riskless firm. While Proposition IV(1) is a consequence of our stochastic formulation with uncertainty entering only through $w(\theta)$, Proposition IV(2) and IV(3) can be seen to hold more generally.

Homogeneous expectations also permit a further delineation of our concept of fair return. Recall that a fair return was defined as a pattern $\Pi^*(\theta)$ of profits across states of nature such that $V = K^0$. With identical expectations, we can talk about "the" probability of each state of nature, and therefore, about an expected profit $E(\Pi^*)$ which yields $V = K^0$. The ratio $r^* = E(\Pi^*)/K^0$ could be called an "expected fair return." But, a regulatory agency would be ill-advised to regulate by limiting expected return to r^*. Examination of first-order "stockholders' interests" conditions subject to the rate of return constraint indicates that neither technical efficiency nor maximal output will be achieved.[30]

---

with other assets held in the portfolio. Therefore, $E(\Pi)/V > r$ is a necessary condition for V to clear the market for the firm's stock. Thus equilibrium for the risky firm A implies $E(\Pi^A)/V = E(\Pi^A)/K^0 > r$, or $E(\Pi^A) > rK^0 = E(\Pi^B)$.

[30] First-order conditions with respect to K^b give

$$E_i\{U_i'(R_v, \theta)[-r - w(\theta)L_K]\} + \lambda[-r - \overline{w}L_K] = 0,$$

where λ is the Lagrangian multiplier associated with the rate of return constraint $E(\Pi)/K^0 \leqq r^*$. If $\overline{w}(\theta) = \overline{w}$ with probability one (certainty), then the above equation reduces to

$$E_i\{U_i'(R_v, \theta) + \lambda\}[-r - \overline{w}L_K] = 0$$

or

$$-L_K = r/w,$$

the efficiency condition. But when $w(\theta)$ is random, it cannot be factored out, and the technical condition (10) will not hold.

6. Some real and imagined problems

■ **Incentives.** If price is adjusted by the regulatory agency to equate market value V with asset value K^0, will there be incentive for the firm to adopt improved techniques? A similar problem arises in the theory of competition with free entry: if new techniques lead to immediate adoption by others and to entry, instantaneously driving down the return to the previous fair rate, no firm will introduce new techniques. But if entry is slow, allowing temporary excess returns, innovation will take place. As suggested by Myers,[31] a similar "conscious lag" in regulation could be used to imitate the competitive solution, and thus provide the impetus for innovation.

□ **Dynamics.** Our model can readily be interpreted in a multiperiod framework. Profit is the discounted present value of the profit stream, dependent on the firm's decisions and the state of nature, and the regulated price p is a *sequence* of prices over time (which perhaps will be revised by future regulatory proceedings).[32] Even though current earnings might be very low, the firm may still be earning a fair return on its capital ($V = K^0$) because of investors' expectations of future profits, given the sequence of regulated prices. Therefore, estimates of fair return based on earnings-price ratios, such as the proposed method of Gelhaus and Wilson, may result in highly "incorrect" regulatory policies for long periods of time.[33]

□ **Volatility of stock prices.** Volatile stock prices could lead to rapidly changing price ceilings if a regulatory agency seeks to equate V and K^0. Yet the very realization that the agency will act in this manner should serve to reduce short-run fluctuations of the firm's stock market value. Even so, if one presumes that stock movements reflect undesirable "noise" as well as underlying trends, the regulatory agency presumably should look not just at current stock market prices, but at some weighted average of stock market value over time. This averaging procedure would tend to reduce the effect of random movements about the stock's intrinsic equilibrium value.

□ **Measurement of K^0.** Our model has been based on the assumption that capital assets can be freely bought or sold at a market price normalized to one. Yet in fact this is rarely an accurate description. Capital is not a homogeneous good, and there are not markets for every vintage of every machine in every location.

---

[31] In [6].

[32] The appropriate discount rate is the certainty interest rate r. Of course, the "certainty equivalent" of present value will be lower, the more uncertainty the investor perceives.

[33] Gelhaus and Wilson [3] suggest allowing a return at time t on book value (for comparability, K^0) equal to E_t/P_t where E_t is current earnings per share and P_t is price per share. That is, regulatory authorities are instructed to set E (E_{t+1}), the expectation of earnings in the next period, equal to $(E_t/P_t)K_0$. Even under certainty, there is no assurance that this policy will assure $\lim_{t \to \infty} V_t = K_0$—it depends on properties of P_{t+1} as a function of E_{t+1} (or earlier earnings). (Note $V_t = P_t S_t$, where $S_t =$ number of shares at time t.)

Buying and selling prices are likely to differ because of installations costs, etc. And some assets, such as "goodwill," may be genuine in the sense of having true economic values, although rarely included in book value or other proximate measures of K^0. In some regulated industries, the measurement problem is further exacerbated by the fact that the price of the capital good depends upon the amount the industry uses.

But the same problem of estimating the economic value of assets currently confronts the regulatory agencies. Present rate determination proceedings require not only an estimation of the "fair" rate of return, but also the "rate base" to which the rate of return is applied. Our theory eliminates the problem of defining fair return, but the measurement problem associated with the rate base remains.

7. Conclusions

■ The model of production and financial equilibrium developed in our Symposium paper[34] has provided a basis for a theory of optimal regulation, when policies must guarantee a fair rate of return to investment. Central to our theory is an implementable definition of "fair return" in the presence of uncertainty and nonhomogeneous expectations.

A fair return was defined as a pattern of returns across states of nature just sufficient to equate the financial value of the firm with the value of its real assets. Correct measurement of real asset value or "rate base," a problem which currently vexes regulatory proceedings, remains a practical hurdle to the implementation of our theory.

Our principal result is that regulating price to equate financial value with real asset value will achieve optimal resource use. Optimality was defined as maximal output subject to (1) an allocationally efficient use of inputs, and (2) a fair return to investors. We further showed that a regulatory agency should not place a ceiling directly upon the expected rate of return: this mode of regulation has undesirable effects on resource allocation.

References

1. AVERCH, H. AND JOHNSON, L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. BAUMOL, W. AND KLEVORICK, A. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
3. GELHAUS, R. AND WILSON, G. "An Earnings-Price Approach to Fair Rate of Return in Regulated Industries." *Stanford Law Review*, Vol. 20, No. 287 (1968).
4. KLEVORICK, A. "The Behavior of a Firm Subject to Stochastic Regulatory Review." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 57–88.
5. LELAND, H. "Production Theory and the Stock Market." *The Bell*

---

[34] See [5].

Journal of Economics and Management Science, Vol. 5, No. 1 (Spring 1974), pp. 125–144.

6. MYERS, S. "The Application of Finance Theory to Public Utility Rate Cases." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 58–97.
7. SCITOVSKY, T. *Welfare and Competition*, 2nd ed. Homewood, Ill.: Richard D. Irwin, Inc., 1971.
8. ZAJAC, E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117–125.

A theory of interdependent demand for a communications service

Jeffrey Rohlfs

Bell Laboratories
Murray Hill, New Jersey

The utility that a subscriber derives from a communications service increases as others join the system. This is a classic case of external economies in consumption and has fundamental importance for the economic analysis of the communications industry. This paper analyzes the economic theory of this kind of interdependent demand. We begin by defining "equilibrium user set" as a set of users consistent with all individuals' (users and nonusers) maximizing their utilities. There are typically multiple equilibria at any given price, and which equilibrium is attained depends partly on the static model, partly on the initial disequilibrium conditions, and partly on the disequilibrium adjustment process. Some general properties of equilibrium user sets are derived. Then we turn our attention to some specific models based on simple characterizations of communities of interest. The implications for pricing are discussed, with special reference to the problem of starting up a new communications service (e.g., a video communications service).

1. Introduction

■ The utility that a subscriber derives from a communications service increases as others join the system. This is a classic case of external economies in consumption and has fundamental importance for the economic analysis of the communications industry. It suggests that although marginal cost pricing may be superior to allocated-cost formulae, it is still not completely appropriate.

This can be illustrated with respect to an historical policy of the industry: promoting universal service. This policy might be justified on the basis of marginal cost pricing, so long as new subscribers pay the incremental cost of expanding the system to accommodate them—even if they do not pay their "allocated"

Jeffrey Rohlfs received the A.B. in economics from Amherst College (1965) and the Ph.D. in economics from M.I.T. (1969). Currently, he is a Member of the Technical Staff at Bell Labs, where his research interests include applied microeconomics, micro theory, and econometrics.

In the course of this study, I talked to many people, and their knowledge and ideas contributed to much of the analysis in this paper. M. Wish has been a collaborator in some previous related work, and he has greatly influenced my thinking. My analysis of the general problem was greatly stimulated by a discussion I had with E. Gilbert, who developed some preliminary results about the maximum equilibrium set. I have also had many profitable discussions with W. Ballamy, J. Berrier, A. Ciesielka, D. Deutsch, A. Gersho, E. Goldstein, A. H. McKeage, D. Mitra, and R. Sanders. F. Sinden, W. Taylor, N. Valcoff, M. Wilk, and E. Zajac provided helpful comments on previous written and oral presentations of this material.

share of average costs. A still lower price, perhaps much lower, might be justified if the externalities are taken into account. The total benefits that *all* subscribers derive from the expansion of the service may be sufficient to justify the incremental costs—even if the new subscribers are unwilling to pay the entire incremental costs.

Recently Artle and Averous[1] made what appears to be the first published analysis of these externalities in communications.[2] They formulate a simple model in which the incremental utility of the service to an individual depends only on the number of telephone subscribers—not on who they are. This is the uniform calling model discussed in Section 3 of this paper. They also assume that the cost of providing telephone service depends only on the number of subscribers. This enables them to derive and interpret the necessary conditions for a social welfare optimum. Their expression has some important similarities (but also some differences) with the usual necessary conditions for a social optimum with respect to a pure public good.

The authors then use these notions to develop a dynamic demand model. They show that interdependent demand can sustain continual growth in a stationary population with stationary income. The mechanism is as follows. New subscribers join. This increases the incremental utility of the service and induces marginal nonusers to join. That in turn induces further growth, etc., etc. The authors offer this as a possible explanation for the continual growth of telephone service observed in all empirical studies of the industry.

Squire studies the problem using a somewhat different model.[3] He considers usage of the system as well as number of telephones, and assumes that the cost of providing the service is a function of these two variables. Squire specifies individual demand curves (based on a fixed number of subscribers) for incoming and outgoing calls. This enables him to develop an expression for optimal usage of the system, based on a modified consumer-surplus concept. He then derives the optimal price per call (charged only to the person making the call) consistent with this optimal usage. He finally develops an expression for the optimal size of the system and the price per telephone consistent with this optimum.

This paper makes a much more detailed analysis of the demand side of the market than attempted by Artle and Averous or by Squire. We begin by defining an "equilibrium user set" as a set of users consistent with all individuals' (users and nonusers) maximizing their utilities. A basic result is that there are typically multiple equilibria at any given price. For example, a very small equilibrium user set may be consistent with utility maximization, since the smallness of the user set in itself makes the service relatively unattractive to potential users. However, a much larger user set may also be possible for the same population at the same price. In this case the largeness of the user set would make the service attractive and allow a high level of demand to be sustained. In any planning (public or private) for the communications service, special attention must be paid to which equilibrium user set is likely to be attained.

---

[1] See [1].

[2] An earlier attempt to model interdependent demand is Marris [2]. Marris develops a general theory of demand for new products but does not consider communications in particular. The interdependent aspects of demand have a much different interpretation in his analysis than in this paper or in Artle and Averous [1] or Squire [3] (discussed later in the text). Also, irreversibility plays a much larger role in Marris' analysis. Nevertheless, Marris develops some of the same concepts used in this paper; e.g., critical mass.

[3] In [3].

The next section of this paper develops a general theory of demand. It derives the following results:

(1) The static model determines the attained equilibrium user set (at a given price) within a certain set of bounds.
(2) A possibly narrower set of bounds (for a given price) is derived, given the initial user set.
(3) Within the bounds defined in (2), the equilibrium attained (at a given price) depends entirely on the disequilibrium adjustment process.

The following two sections of the paper develop specialized models based on various simple characterizations of communities of interest. The simplest of all is the uniform calling pattern, which assumes that no one has any special community of interest (other than the entire population). This model is the only one in which the equilibrium theory can be developed in terms of the *number* of users, without paying attention to who they are. We can therefore define a demand curve, which turns out to have an inverted *U* shape. See Figure 1 on page 28.

Zero demand is a stable equilibrium for all positive prices. The upward-sloping part of the inverted *U* consists of unstable equilibria and constitutes the "critical mass" of the service (at any given price). If the critical mass is exceeded, demand expands to the downward sloping part of the inverted *U*. Points on the latter are stable equilibria and represent the maximum level of demand sustainable at a given price.

Unfortunately (for ease of analysis), the uniform calling pattern may not be very realistic. People typically belong to groups, each of which has a strong community of interest within itself. And they typically have a few principal contacts who alone account for a substantial part of their communication. These complications are briefly discussed in the section entitled "Nonuniform Calling Patterns."

The final section of the paper discusses some implications of the preceding demand analysis for supply and pricing of the service. An important distinction is made between viability of the service (existence of a nonnull equilibrium user set that can be served with nonnegative profits) and the start-up problem (how to attain such a user set, starting from a small or null initial user set).

Viable nonnull equilibrium user sets (if they exist) are always superior to the null set from a static point of view. We can compare such sets to determine the static social optimum or the overall market equilibrium corresponding to a static supply model. However, this kind of analysis is incomplete and may be misleading without consideration of the start-up problem. Achieving the static optimal user set may require ruinous (albeit temporary) promotional costs.

Appropriate solutions to the start-up problem depend in large part on the demand model. In the uniform calling model, the start-up problem is simply a question of getting beyond the critical mass. Community of interest groups may make the practical start-up problem much easier, but they also introduce some special problems. If an individual's demand is contingent on a few principal contacts' being users, there may exist many small self-sufficient user sets. These allow the possibility of a long-term introductory program, in which the seller gradually expands the size and number of such sets.

This paper presents only a limited discussion of costs and supply. The reason is that costs of a communications service are very complex and merit a separate study in their own right. This is a very fruitful topic for future research.

2. General theory of demand

■ Let the population consist of n individuals. As in Artle and Averous' work, we define a set of binary variables:

$$q_i = \begin{cases} 0 & \text{if individual } i \text{ does not subscribe to the communication service} \\ 1 & \text{if the individual } i \text{ does subscribe to the communication service} \end{cases} \qquad (1)$$

for $i = 1, \ldots, n$.

We assume there are also m other goods in the economy. To model interdependent demand, we specify a pair of utility functions for each individual:

$$U_i^0 = U_i^0\,(r_{i1}, \ldots, r_{im}) \qquad (2)$$

$$U_i^1 = U_i^1\,(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, r_{i1}, \ldots, r_{im}) \qquad (3)$$

where

U_i^0 = Utility of individual i if he does not subscribe to the communications service,

U_i^1 = Utility of individual i if he does subscribe to the communications service, and

r_{ij} = Consumption of (noncommunications) good j by individual i.

Equations (2) and (3) implicitly assume independent utilities with respect to all goods in the economy *other* than the communications service in question. In addition, we make the usual monotonicity assumptions:

$$\frac{\partial U_i^k}{\partial r_{ij}} \geqslant 0 \text{ for all } j \text{ and } > 0 \text{ for some } j; \quad \text{and} \qquad (4)$$

$$U_i^0 \leqslant U_i^1 \qquad (5)$$

for all $i, k, q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, r_{i1}, \ldots, r_{im}$.

We also make a specialized assumption applicable to a communications service:

$$\frac{\partial U_i^1}{\partial q_w} \geqslant 0 \qquad (6)$$

for all $i \neq w$, $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, r_{i1} \ldots, r_{im}$. That is, a subscriber's utility never decreases as additional individuals subscribe (and none drop out).

This seems like a reasonable working assumption. We can, of course, imagine some exceptions; e.g., the value of the service to others would probably be lessened if a large number of life insurance salesmen subscribed to the service to solicit other subscribers. However, we assume that such occurrences are the exception rather than the rule—that, in general, the availability of a communications link is not detrimental to either party.

We assume utility maximization, which we analyze in two steps. (1) We evaluate the maxima of U_i^0 and U_i^1 (with respect to $r_{i1}, \ldots, r_{im}$) subject to individual i's budget constraint. Let us denote these maxima as $\hat{U}_i^0$ and $\hat{U}_i^1$. (2) We then compare $\hat{U}_i^0$ and $\hat{U}_i^1$ to see if the individual demands the communication service. This defines a demand variable for each individual:

$$q_i^D = \left\{ \begin{array}{l} 0 \text{ if } \hat{U}_i^0 > \hat{U}_i^1 \\ 1 \text{ if } \hat{U}_i^0 \leqslant \hat{U}_i^1 \end{array} \right\} \qquad (7)$$

for $i = 1, \ldots, n$.

The basic methodology of this paper is to ignore interrelationships between the communications market and other markets and concentrate on relationships *within* the communications market. Thus, we make the *ceteris paribus* assumption that prices of all goods other than the communications service are fixed and that each individual has a fixed budget constraint. This allows us to express the demand variables as functions of price and the set of subscribers:

$$q_i^D = q_i^D\ (p, q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n) \qquad (8)$$

for $i = 1, \ldots, n$, where $p =$ the price of the communications service.

It follows from previous assumptions that all the q_i^D are monotonically decreasing (equality allowed) with respect to p. That is, an increase in p can never change q_i^D from 0 to 1; a decrease in p can never change q_i^D from 1 to 0. However, a change in p may have no effect on q_i^D. It also follows from previous assumptions that all the q_i^D are monotonically increasing (equality allowed) with respect to all q_w $(w \neq i)$.

☐ **Equilibrium user sets.** Naturally, there is a correspondence between demanding the service and being a subscriber. We define an *equilibrium user set* as a set of users such that

$$q_i = q_i^D\ (p, q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n) \qquad (9)$$

for all i.[4] Thus, in equilibrium all users demand the service; all nonusers do not demand it.

Equation (9) defines equilibrium with respect only to the demand side of the market. It describes user sets that are consistent with utility maximization at a given price. These constitute necessary but not sufficient conditions for an overall market equilibrium. The latter additionally requires that the user set and price be consistent with some specified model of supply behavior.

For fixed $p = \bar{p}$, equations (9) are a system of n equations in n binary variables. Such a system does *not* generally have a unique solution. In fact, unique solutions did not arise in any of the simple models investigated in this paper (except in the trivial case where price is so high that there can never be any demand at all).

Consequently, the equation

$$q = q^D, \tag{10}$$

where

$$q = \sum_{i=1}^{n} q_i \quad \text{and}$$

$$q^D = \sum_{i=1}^{n} q_i^D$$

may be indeterminate (for fixed $\bar{p}$). That is, it may either hold or fail to hold depending on which set of users constitutes the sum q.

For this reason, the general theory of interdependent demand cannot be developed in terms of the sum q. It is necessary to work with the individual q_i. The basic analytical concept is not the demand curve—i.e., equilibrium pairs (q, p)—but rather equilibrium user sets.

☐ **Disequilibrium analysis.** Given that several equilibrium user sets exist for a given price (*ceteris paribus*), it is important to know which ones (if any) are most likely to occur. This requires analyzing what happens if the market is initially in disequilibrium. Our procedure is as follows. We specify a very general disequilibrium adjustment process. We then investigate the extent to which the user sets resulting from this process depend on the static model, the extent to which they depend on the initial disequilibrium conditions, and the extent to which they are indeterminate, depending on a more detailed specification of the adjustment process.

In this section, we restrict our attention to the demand side of the market and assume a given price for the communications service. We further assume that adjustments of consumption in other markets can be made rapidly and costlessly. This seems like a reasonable simplifying assumption, allowing us to analyze dis-

---

[4] We can also define equilibrium user sets with respect to any given set of discriminatory prices; i.e., a set of users such that

$$q_i = q_i^D(p_i, q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n) \tag{9a}$$

for all i, where p_i is the price charged to individual i.

equilibria in the communications market without considering possible disequilibria in the rest of the economy.

Now suppose there is an arbitrary initial user set. It may be based on utility maximization for current or previous states of the world, past selling efforts of the supplier of the service, or anything else. We assume that adjustments to this user set occur according to the following adjustment process. (1) An individual in equilibrium $(q_i^D = q_i)$ never changes his status from user to nonuser or vice versa. This is reasonable, since such a change would always reduce his utility (except in the knife-edge case where $\hat{U}_i^0 = \hat{U}_i^1$, in which case the change in status has no effect on utility). (2) The length of time an individual can remain continually in disequilibrium $(q_i^D \neq q_i)$ is bounded. He eventually must change his status. This is also reasonable, since the change always increases his utility (except in the knife-edge case where $\hat{U}_i^0 = \hat{U}_i^1$, in which case the change in status has no effect on utility).

The adjustment process is essentially a model of utility maximization with inertia. It is very general in that it makes no assumption about the speed of adjustment. This speed may vary from individual to individual. It may depend on the user set or actions of the seller. Or it may change over time.

A limitation of this process is that it does not allow individuals to collude and subscribe together. This is relatively unimportant if an individual's demand is contingent on a large user set, since such collusion would be difficult with very large groups. However (as will be seen later) we do have to consider relaxing the assumption in models where an individual's demand is contingent on a few of his principal contacts' being users.

It is important to note that the adjustment process does *not* necessarily converge to an equilibrium user set. Consider the following example. *A* demands the service if and only if *B* is a user; *B* demands the service if and only if *C* is a user; *C* demands the service if and only if *A* is a user. Suppose the initial user set is *A*. One possible version of the adjustment process is as follows. *C* joins because *A* is a user. Then *A* disconnects because *B* is not a user. Then *B* joins because *C* is a user. Then *C* disconnects because *A* is not a user. Then *A* joins because *B* is a user. Then *B* disconnects because *C* is not a user. We are now back to the original user set, and the process can be repeated indefinitely.

Nevertheless, the user sets resulting from the adjustment process can be bounded, as shown in the following theorems.

Theorem 1: If the initial user set is the entire population, the adjustment process can only *remove* individuals from the user set; no individual can ever be added who has previously dropped out.

Proof: If the entire population is an equilibrium user set, no one is added or removed, and the theorem is satisfied. If the entire population is *not* an equilibrium set, let $r_1, r_2 \ldots$ represent the sequence of individuals who change status. (If individuals change status simultaneously, we list them in arbitrary order.) Now r_1 must be a removal (not an addition), since the entire population consists of users, and there is no one left to be added. Given that $r_1, \ldots, r_k$ are removals, r_{k+1} must also be a removal for the following reason. The only possible additions would be the individuals $r_1, \ldots, r_k$. But all of these dropped out (and therefore did not demand the service) when the user set contained the current user set. Thus, they cannot demand the service according to the monotonicity assumption.

It follows that all of the r_i must be removals. Q.E.D.

Theorem 2: If the initial user set is the entire population, the adjustment process converges to an equilibrium user set in finite time.

Proof: The process must converge in finite time for the following reason. All changes of status are removals. Since only n (the size of the population) individuals can be removed, there are at most n changes in status. These must all occur in finite time.

After all the changes in status occur, no user can fail to demand the service, for otherwise the process would continue. Moreover, no nonuser can demand the service because of the monotonicity condition. Thus, the final user set is an equilibrium user set. Q.E.D.

Theorem 3: If the initial user set is the entire population, the adjustment process converges to the union of all equilibrium user sets regardless of the order of removals.

Proof: Let X be an arbitrary equilibrium user set; let R be the equilibrium result after individuals $r_1, \ldots, r_k$ have been removed according to the adjustment process. X cannot contain r_1, since r_1 was removed when the user set was the entire population (and hence contained X). Since X does not contain r_1, it cannot contain r_2 either. (The entire population minus r_1 contains X-r_1.) Nor can X contain $r_3, \ldots, r_k$. Thus, $X \subseteq R$.

But X is an arbitrary equilibrium user set. Thus, R contains all equilibrium user sets. Since R is itself an equilibrium user set, it is the union of all equilibrium user sets. Q.E.D.

This set will hereafter be referred to as the "maximum equilibrium user set."

By entirely symmetrical reasoning we can establish the following. If the initial user set is null,

(1) the adjustment process can only add individuals, and no one is ever removed who previously joined;
(2) the adjustment process converges to an equilibrium user set in finite time;
(3) the adjustment process converges to the intersection of all equilibrium user sets, regardless of the order in which individuals are added.

This set will hereafter be referred to as the "minimum equilibrium user set."

It is important to note that the minimum and maximum equilibrium user sets need not be the same. For example, consider the following simple model. Each individual demands the service if three of his five principal contacts are users. The minimum equilibrium user set is null. There are no users; so no one has three of his five principal contacts as users; so no one demands the service. The maximum equilibrium set is the entire population. Everyone has all five of his principal contacts as users; so everyone demands the service. In addition, there may be any number of equilibrium user sets between these two extremes, depending on the distribution of principal contacts.

In this example, the static model tells us practically nothing about what equilibrium will actually be attained. We only know that it will be zero or 100 percent or somewhere in between. This is an extreme case, but in general the static model determines the actual equilibrium only within certain bounds—the minimum and maximum equilibrium sets. Moreover, in all of the models investigated in this paper, there exists the possibility that these bounds may be far apart. In a practical situation, this difference may mean the difference between marketing success and failure.

The above theorems show that if the initial user set is sufficiently large, convergence to the maximum equilibrium user set is assured (according to the assumed adjustment process). If the initial user set is sufficiently small, convergence to the minimum equilibrium set is assured. For intermediate initial disequilibrium user sets, the actual equilibrium attained may also depend on a more detailed specification of the adjustment process than given above.[5] It may be critical whether or not the disequilibrium nonusers subscribe before the disequilibrium users drop out.

☐ **Particular initial user set.** The minimum and maximum equilibrium user sets provide bounds on user sets that are possible for *any* given initial user for *any* version of the adjustment process described above.[6] This subsection provides bounds on user sets that can be attained from a *particular* initial user set for *any* version of the adjustment process.

Let S be an arbitrary initial user set. We now define the following two adjustment sequences.

Optimistic sequence

(1) First all nonusers who demand the service subscribe in arbitrary order, but no users drop out. This converges to the same user set $\bar{\bar{S}}$, irrespective of the order in which individuals subscribe. (Proof is analogous to that of Theorem 3.)

(2) Then all users who do not demand the service drop out in arbitrary order. This converges to the same user set S, irrespective of the order in which individuals drop out. Moreover, $\bar{S}$ is an equilibrium user set. (Proof analogous to Theorems 2 and 3.)

Pessimistic sequence

(1) First all users who do not demand the service drop out in arbitrary order, but no nonusers subscribe. This converges to the same user set $\underline{\underline{S}}$, irrespective of the order in

---

[5] P. 22.

[6] P. 22.

which individuals drop out. (Proof analogous to Theorem 3.)

(2) Then all nonusers who demand the service subscribe (in arbitrary order). This converges to the same user set $\underline{S}$, irrespective of the order in which individuals subscribe. Moreover, $\underline{S}$ is an equilibrium user set. (Proof analogous to Theorems 2 and 3.)

Now let $R_1, R_2, \dots$ be a sequence of user sets resulting from applying an arbitrary version of the adjustment process to S. As previously discussed, this sequence need not converge to an equilibrium user set. However, we can place the following bounds on the sequence:

(1) $\underline{\underline{S}} \subseteq R_i \subseteq \overline{\overline{S}}$ for all i. This follows directly from the monotonicity assumption.

(2) After some finite period of time, $\underline{S} \subseteq R_i \subseteq \overline{S}$ for all i.

Proof: Let $x_1, \dots, x_k$ be a sequence of individuals who drop out in part (2) of the optimistic sequence. Since x_1 does not demand the service given user set $\overline{\overline{S}}$, he cannot demand it given any user set R_i. Thus, if x_1 is a user, he is continually in disequilibrium. He must drop out in finite time and can never thereafter rejoin. Once x_1 drops out, we can apply the same reasoning sequentially to $x_2, \dots, x_m$. Thus, after some finite period of time, all $R_i \subseteq \overline{S}$. The proof that $\underline{S} \subseteq R_i$ is exactly symmetrical. Q.E.D.

Thus, the optimistic and pessimistic sequences define bounds on user sets attainable from a particular initial user set. These bounds may (or may not) be considerably narrower than the bounds provided by the minimum and maximum equilibrium sets. In any event, within these bounds the equilibrium user set attained depends entirely on a detailed specification of the adjustment process.

☐ **Additive utilities.** In order to proceed further we must make more assumptions. To simplify the problem we propose a model of additive utilities. That is, we define a vector $\underset{\sim}{f}$ and a matrix $\underset{\sim}{V}$ such that

$$U_i^0 = f_i\,(r_{i1}, \dots, r_{im}) \tag{11}$$

$$U_i^1 = f_i\,(r_{i1}, \dots, r_{im}) + \sum_{j \neq i} v_{ij}\, q_j, \tag{12}$$

where v_{ij} $(i \neq j)$ is the incremental utility to individual i of a communications link with individual j, $(v_{ij} \geqslant 0.)$

The additive model assumes that these incremental utilities do not depend on consumption of other goods or on other communications links available to the individual. These do seem to be reasonable simplifying assumptions, but there are some problems with them. The growth of telephone service has had fundamental effects on social and business customs, and these would not be captured in an additive model. It has also resulted in substantial changes in communities of interest, which are assumed to be fixed in equation (12). However, the additive model would be commensurately better for analyzing smaller differences in market penetration or for analyzing a service that does not provide so revolutionary an improvement in communications as did the invention of the telephone.

Equation (12) also assumes that the service has no value except to communicate with others who have the service. The service is worthless if no one else subscribes. This assures that the null set is an equilibrium user set at any positive price.

This assumption sounds reasonable enough, but there are some possible exceptions. An individual may have noncommunications applications for the hardware. If the service is new, he may find it prestigious or derive self-satisfaction from being an innovator. However, these kinds of considerations go beyond the scope of this paper.

The additivity assumption is quite useful and allows us to derive a convenient expression for q_i^D as shown below.

The maxima $\hat{U}_i^0$ are defined by the *ceteris paribus* conditions and do not depend on anything in the communications industry. Maximizing equation (12) with respect to $r_{i1}, \ldots, r_{im}$, subject to individual i's budget constraint, we obtain

$$\hat{U}_i^1 = \hat{U}_i^0 - h_i(p) + \sum_{j \neq i} v_{ij} q_j \tag{13}$$

for some function h_i where $h_i(0) = 0$, $h_i(p) > 0$ for all i.

It follows that

$$q_i^D = \begin{Bmatrix} 0 \text{ if } \sum_{j \neq i} v_{ij} q_i < h_i(p) \\ 1 \text{ if } \sum_{j \neq i} v_{ij} q_j \geqslant h_i(p) \end{Bmatrix} \tag{14}$$

where $v_{ij} \geqslant 0$ for all i, j.

We also assume constant marginal utility of money for a given individual. This means that $h_i(p)$ is a linear function:

$$h_i(p) = b_i p. \tag{15}$$

We can therefore write equation (14) as follows:

$$q_i^D = \begin{Bmatrix} 0 \text{ if } \sum_{j \neq i} w_{ij} q_j < p \\ 1 \text{ if } \sum_{j \neq i} w_{ij} q_j \geqslant p \end{Bmatrix} \tag{16}$$

where $w_{ij} = \dfrac{v_{ij}}{b_i}$ for all $i \neq j$.

☐ **Further simplification.** Both the monotonicity and the additivity assumptions greatly simplify the problem (at some cost in realism). However, we still must deal with the matrix $\underset{\sim}{V}$ which is the size of the population squared. Thus for a city with a population of one million, $\underset{\sim}{V}$ would have one trillion entries. Clearly, further simplification is required. The following two sections consider some possibilities for breaking the problem down to manageable size.

3. Uniform calling pattern

■ The preceding section began by considering the problem in its full generality and considered some reasonable kinds of simplifying assumptions. We now take the opposite approach, beginning with a very simple model and then relaxing assumptions to make the model more complicated and realistic.

In this section we assume that all the (off-diagonal) elements in any single row of $\underset{\sim}{V}$ are equal. This implies that no one has any special community of interest other than the entire population. The *number* of subscribers affects an individual's demand, but he does not care who these subscribers are.

This may in fact be a reasonable approximation (for some purposes). We might reason that the individual communicates with a large number of people during the course of a year, many of whom he does not know in advance. The number of users may be as good a proxy as any for the incremental utility he derives from the service.

However, it is also true that most people belong to groups, each of which has a community of interest within itself. They also typically have a few principal contacts with whom they communicate more than with others. Thus, their demand for a communications service would depend on how many members of their community of interest group and which of their principal contacts subscribe to the service.

In any event, the uniform calling model seems like a good place to begin developing the theory. (This model is also adopted by Artle and Averous and Squire.)[7] It allows some strong results to be derived and provides some useful insights about interdependent demand. Results from the uniform calling model also provide convenient reference points for analyzing more complex models, which are briefly discussed in the next section of this paper.

The uniform calling model allows us to write equation (15) as follows for a large population:

$$q_i^p \left\{ \begin{matrix} 0 \text{ if } fw_i < p \\ 1 \text{ if } fw_i \geqslant p \end{matrix} \right\} , \qquad (17)$$

where $f =$ the user fraction (q/n), and

$$w_i = \sum_{j \neq i} w_{ij}.$$

This in turn allows individuals to be ordered in terms of their demand for the service. That is, if $w_i \geqslant w_j$ (i's demand exceeds j's), individual i is in every equilibrium user set that contains j.

□ **Demand curve.** Since individuals can be ordered as above, every equilibrium user set consists of *all* individuals (i) for whom $w_i \geqslant$ some K. Similarly, for any q, there is at most one equilibrium user set with q members; i.e., the q people with the highest values of w_i. (If more than one person has the minimum w_i in the user set, all persons with that w_i must be in the user set for it to be an equilibrium.)

---

[7] In [1] and [3], respectively.

Thus, every equilibrium user set can be uniquely characterized by q, the number of members in it. We can develop the equilibrium theory for this model in terms of the sum q, without specifying the individual elements q_i. In particular, we can define a demand curve; i.e., the locus of all the pairs (q, p) for which there exists an equilibrium user set (which would have to be the q people with highest w_i).

This gives us a convenient way of looking at the relationship between price and the equilibrium user sets. However, it is important to note that equation (10) is still indeterminate, and we must be careful in applying the demand curve in disequilibrium situations.

☐ **An example.** Before discussing the general properties of such a demand curve, let us consider a specific example. Suppose the population is large, and w_i is distributed uniformly between 0 and 100 over the population. For the marginal individual

$$w_i = 100(1 - f). \tag{18}$$

For an equilibrium at $0 < f < 1$, fw_i for the marginal individual must equal p. [See equation (17).] Thus, the demand curve is the locus of points where

$$100\,f(1 - f) = p. \tag{19}$$

As previously mentioned[8] the null set ($f = 0$) is an equilibrium user set for all $p > 0$. For an equilibrium at $f = 1$, p must be less than or equal to fw_i for all individuals. But the minimum of w_i and hence fw_i is 0. Thus, the only equilibrium is $p = 0$.

FIGURE 1

DEMAND CURVE FOR UNIFORM CALLING PATTERN

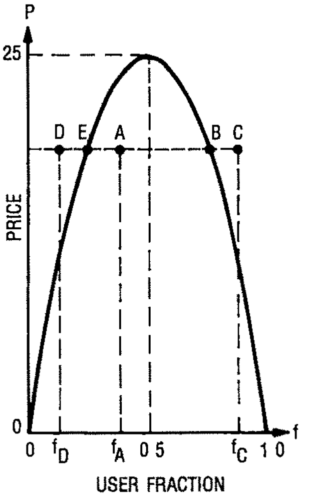


Figure 1 shows the demand curve. It consists of the entire positive p-axis plus an invested parabola going through (0,0) and (1,0) and having a maximum at (0.50,25).

The maximum equilibrium set is the right-hand side of the parabola for $0 < p \leqslant 25$; it is null for $p > 25$.

For small p, there is an enormous difference between the minimum and maximum equilibrium user sets. Thus, the actual equilibrium attained (for small p) depends critically on the initial disequilibrium conditions and the disequilibrium adjustment process.

☐ **Disequilibrium analysis.** The following analysis of disequilibrium is based on the adjustment process proposed above.[9] As before, we assume a fixed price and restrict our attention to the demand side of the market. We first consider the special case in which the initial disequilibrium users are those with the highest w_i. We then consider the general case by examining arbitrary perturbations from an initial equilibrium.

☐ **Initial users have highest w_i.** This subsection assumes that the initial user set consists of all individuals for whom $w_i \geqslant$ some K. This is necessarily true if the users form an equilibrium user set for any $\bar{p}$. Thus, the results apply to any disequilibrium brought about by a price change from an initial equilibrium.

---

[8] P. 26.

[9] P. 22.

Suppose we are originally in disequilibrium at A (in Figure 1), underneath the parabola. Given the user fraction f_A, the equilibrium price is higher than the actual price. All users are satisfied, but some nonusers would prefer to become users. If p remains constant, the user fraction will ultimately increase to B.

Suppose we are originally in disequilibrium at C. Given f_C, actual price exceeds equilibrium price, and f declines to B.

Suppose we are originally in disequilibrium at D. Given f_D, actual price exceeds the equilibrium price. So, f declines. As f declines, the discrepancy between actual and equilibrium price increases until the market achieves equilibrium at $f = 0$.

In all these cases, the order in which individuals join or drop out is immaterial. The optimistic and pessimistic sequences are equivalent, and all versions of the adjustment process converge to the same equilibrium user set.

In general, the positive p-axis and the downward sloping part of the parabola in Figure 1 consist of stable equilibria. The upward sloping part of the parabola consists of unstable equilibria.

The upward sloping part of the parabola can be regarded as a "critical mass" for the service. That is, for any positive price below the maximum of the parabola, the market must be forced to some initial disequilibrium beyond the critical mass before the service can grow by itself. The higher the price, the higher is the critical mass.

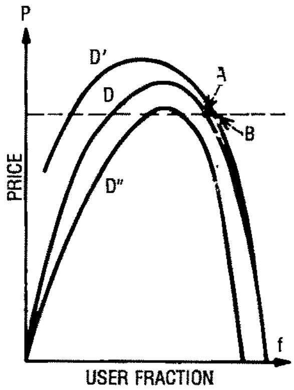
CALCUTTA CENTRAL LIBRARY UNIVERSITY

☐ **Arbitrary perturbations from equilibrium.** The preceding analysis does *not* necessarily apply for arbitrary initial conditions. An initial user set of y people may converge to very different equilibria depending on who those y people are and on a more detailed specification of the adjustment process (than given above[10]).

For example, suppose the initial user set of y people contains none of the y people with highest w_i. Suppose it contains the people with the next y highest values of w_i. Even though the initial market penetration is y, the market can achieve a critical mass of $2y$ if the y nonusers with highest w_i all subscribe before any of the initial users drop out (optimistic sequence). However, if all the initial users who do not demand the service drop out before any nonusers subscribe (pessimistic sequence), the market may fail to achieve a critical mass much lower than y.

FIGURE 2

ARBITRARY PERTURBATIONS FROM EQUILIBRIUM



The results of the preceding subsection do apply for small but arbitrary perturbations (in the user set) from an original stable equilibrium. For example, suppose the market is originally in the stable equilibrium at A in Figure 2, and a set (R) of nonusers subscribes. We can analyze this by constructing a demand curve conditional on all members of R being users (D' in Figure 2). The perturbation of R becoming users may cause additional users to subscribe. However, no matter what the adjustment process is (subject to the rules laid down above[11]), demand can never expand beyond B. And at B (or any point between B and A), only people in the *original* equilibrium set demand the service. Thus, the market eventually goes back to A (so long as the price remains fixed).

---

[10] P. 22.

[11] P. 22.

In the general case the perturbation may involve users dropping out as well as nonusers joining. We analyze this by constructing D' as before and D'', demand conditional on those who drop out being nonusers. The market might converge to $f = 0$, if the perturbation is large and brings demand below critical mass. However, if the perturbation is sufficiently small, D'' will be sufficiently close to D that demand cannot go below critical mass. Thus, the market must return to the initial equilibrium.

This reasoning also applies to the stable equilibria at $f = 0$. If the perturbation is sufficiently small, D' will be sufficiently close to D that critical mass cannot be achieved, and the market must return to the initial equilibrium.

☐ **General properties of the model.** Some properties of the above example apply generally to all uniform calling models. The entire positive p-axis always consists of stable equilibria. The demand curve always has an upward-sloping part, which constitutes the critical mass for the service for initial users sets with maximal w_i. It always has a downward sloping part (perhaps vertical), which consists of stable equilibria. However, both parts need not be unique, and the demand curve may be jagged. This allows the possibility of many stable equilibria for a given price.

4. Nonuniform calling patterns

■ This section considers some models that are more complex than the uniform calling model. Some specific results are presented, but they are naturally not so strong as those of the previous section. Our primary objective is to point out the analytical complexities in such models and suggest some ways of dealing with them.

☐ **Community of interest groups.** Suppose the population consists of k groups ($k \leqslant n$). We assume that an individual has the same community of interest with everyone in the same group. However, this community of interest may be different for different groups. Mathematically, we assume that if individuals j and m are in the same group, $v_{ij} = v_{im}$ for all i.

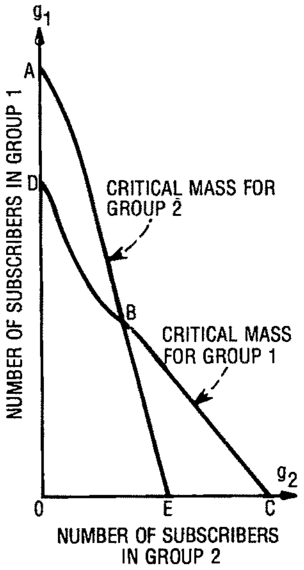
☐ **Disjoint groups.** The simplest case is disjoint groups. That is, $v_{ij} = 0$ unless i and j are in the same group. In such a model, we can consider each group as a separate population, and all the analysis of the previous section carries over. A critical mass can be defined for each group in terms of market penetration within that group. For given $\bar{p}$ the maximum possible number of stable equilibria is 2^k (unless the demand curve for some groups is jagged). The equilibria are characterized by which of the k groups achieve their critical mass.

☐ **Joint groups.** We now consider the case of joint groups, where v_{ij} does not necessarily equal zero for i and j in different groups.

In this model, the incremental utility of the service to an individual is a function of the g_i (number of subscribers in the ith group). Thus, we can define the critical mass for each group as a function of all the g_i (and p). This is illustrated in Figure 3 for the case of two groups.

FIGURE 3

CRITICAL MASSES FOR MODEL WITH COMMUNITIES OF INTEREST



If the initial user set in Figure 3 consists of individuals with the highest w_i in each group (a weaker condition than requiring initial users to have the highest w_i in the population), we have the following result. If all groups achieve their critical mass (initial market penetration outside *ABCO* in Figure 3), the service will expand to the maximum equilibrium set. If no group achieves its critical mass (initial market penetration within *DBEO* in Figure 3), the service will collapse to the minimum equilibrium set. If some groups achieve their critical mass but others do not (initial market penetration within *ABD* or *BCE* in Figure 3), the service may expand to the maximum equilibrium set, collapse to the minimum equilibrium set, or achieve equilibrium somewhere in between. Which of these occurs depends on the parameters of the static model, the initial market penetration in each group, and the disequilibrium adjustment process. An upper bound on the number of stable equilibria is 2^k (unless the demand surface is jagged). Each equilibrium is characterized by which of the k groups achieve their critical mass and which do not.

☐ **Further refinements.** Introducing further refinements into the model is straightforward. As before, we somehow divide the population into k groups such that within each group individuals can be ordered in terms of their demand for the service. For greater realism, we could have a large value for k, but that has the drawback of requiring us to deal with a k-dimensional quantity vector and to contend with the possibility of many different equilibria (for each price).

☐ **Few principal contacts.** An individual's demand may depend primarily on which of his few principal contacts are users. A basic analytical tool for studying such demand is the "self-sufficient" user set; i.e., a set of individuals, each of whom demands the service conditional on all others in the set being users. An equilibrium user set must, of course, be self sufficient, but the converse is not necessarily true. Someone outside the self-sufficient set may demand the service if everyone in the set has it.

All self-sufficient sets necessarily belong to the maximum equilibrium user set. Moreover, if the entire self-sufficient set is contained in the initial disequilibrium user set, then the entire self-sufficient set is necessarily part of the final equilibrium user set.

In any practical problem, we could never hope to have a complete empirical list of principal contacts. The way to proceed in such cases is to specify a probability distribution which indicates (approximately) how likely various configurations of principal contacts are. This leads to some interesting combinatorial analysis but goes beyond the scope of this paper.

5. Some implications for supply and pricing

■ Costs of providing a communications service depend on the constituency of the user set as well as its size. These costs are in themselves fully as complex as interdependent demand. Such complexities go beyond the scope of this paper, which is a study of interdependent *demand.* Nevertheless, it is useful to look at some implications of the preceding sections for supply and pricing of the service. We assume that the service is provided by a (regulated) monopoly and specify a very general cost function:

$$C = C(q_1, \ldots, q_n), \tag{20}$$

where C is strictly monotonic (equality *not* allowed) in all its arguments.

We investigate various kinds of pricing strategies, some involving short-run losses. However, we assume that in (long-run) equilibrium, the monopoly must earn nonnegative profits. Some of the pricing strategies considered involve discriminatory pricing, but we do not assume that perfect discrimination is necessarily possible. Indeed, perfect discrimination would surely not be possible in any realistic situation. Consequently, the nonnegative profit restriction may be inconsistent with Pareto optimality. It is nevertheless consistent with existing real-world institutions.

We can now make a crucial distinction for planning supply of a communications service; i.e., the difference between "viability of the service" and "the start-up problem." Viability is determined solely by the static model. It means that there exists a nonnull equilibrium user set that can be served with nonnegative profits. (We also refer to such a user set as "viable.") The start-up problem is a dynamic consideration. It refers to the costs and practical difficulties of attaining a viable user set, starting from a small or null initial user set.

From a static point of view, any viable user set is superior to the null user set (in the sense that the supplier of the service and all users are at least as well off as before and possibly better off. Nonusers' utilities are unchanged).[12] If there are several viable user sets, they can be compared to determine the social optimum (subject to the nonnegative profit restriction). We can also determine the overall market equilibria consistent with various static supply models.

However, this kind of static analysis is incomplete and may in fact be misleading. We must also consider the dynamic aspects; i.e., the start-up problem. If the initial user set is small or null, the static social optimum may require ruinous (albeit temporary) promotional costs. Thus, all things considered, a smaller user set, or perhaps even the null set, may be superior.

The remainder of this section discusses various possible solutions to the start-up problem. We consider the case of a new service (e.g., a video communications service) and assume that the initial user set is null. This is necessarily an equilibrium user set at all positive prices. Thus, the service, even though viable, cannot get started by itself. It requires some positive action by the seller, probably involving temporary losses. The next three subsections consider some possible start-up strategies in the context of the various demand models we have studied in the previous sections.

---

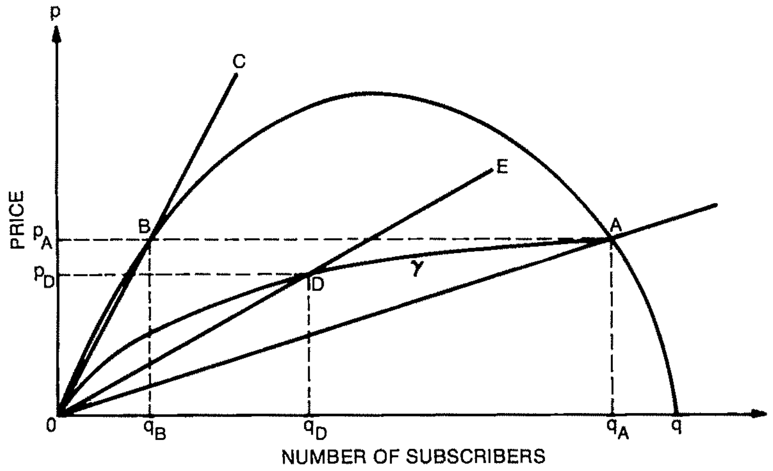
[12] Of course, this issue becomes much more complicated if we consider interrelationships with other markets and possible income redistribution.

Although we assume in this analysis that the product is viable, it is worth noting that in real life the seller would have no such guarantee. There is always a risk involved in introducing a new product or service. The seller generally faces certain losses when the product or service is first introduced and the *prospect* of future profits. In a regulatory environment, the supplier must consider how the regulators will respond if the service succeeds and how they will respond if it fails.

☐ **Uniform calling pattern.** Let us assume that the demand curve is a (nonjagged) inverted U, and suppose the desired nonnull equilibrium set is A in Figure 4. The long-run optimal price is P_A, but some method must be contrived to get beyond the critical mass.

FIGURE 4

SOLVING THE START-UP PROBLEM



☐ **Direct approach.** The most direct approach is to give the service free to a selected group of people for a limited time. For this method to succeed, the initial user set must, of course, be sufficiently large to achieve critical mass. Half measures are worse than useless. If critical mass is not achieved, the whole effort will be a complete failure, and demand will eventually contract to zero.

The success of this approach may also depend on how the initial user set is selected. The discussion above[13] shows that the optimal initial user does not necessarily consist of those with maximal w_i. (In the example, an alternative choice of y people allowed a critical mass of $2y$ to be achieved.) However, if any other set is chosen, there should be some assurance that the initial users will not discontinue the service before those with higher w_i subscribe.

---

[13] P. 29.

If they do discontinue, the start-up effort will have failed to achieve anything.

□ **Low introductory price.** Another way to start up the service is to have a low introductory price. This price could then be raised as the number of subscribers increased. There are any number of ways this could be accomplished. We can represent the introductory program as an expansion path in (q,p) space; e.g., γ in Figure 4. The expansion path shows how p increases as q increases.

The expansion path must pass through the origin to get the service started in this simple model. If, in addition, the expansion path is always concave downward, the introductory program has a very desirable property. Regardless of the order in which individuals enter the market, no individual ever subscribes and then later discontinues service after price and quantity rise along the expansion path. The proof is as follows. Suppose an individual enters at D. That means he is willing to pay p_D to have the service available for q_D of his communication. It follows from the uniform calling assumption that he would be an equilibrium user at *any* point along the straight line OE through D. But the concavity assumption assures that the expansion path γ is always below OE for $p > p_D$, $q > q_D$. Thus, if the individual joins at D, he cannot drop out at any point on γ to the right of D. Q.E.D.

The above condition can be very important if the cost of connecting an individual to the network is large.

Perhaps the most interesting program with a low introductory price is usage-proportional pricing. The remainder of this subsection investigates that plan under various assumptions.

We first consider the possibility that all subscribers have equal usage, proportional to the number of subscribers. It follows that price is also proportional to number of subscribers [instead of proportional to a quadratic function of number of subscribers as in the equilibrium model, equation (19)], and the expansion path of the service is a straight line.

One possibility would be to let $p = p_A \dfrac{q}{q_A}$. In that case demand would expand along OA. Price would increase automatically as q increased. When q reached the optimum, q_A, price would just equal its optimum, p_A.

An alternative is to let $p = p_A \dfrac{q}{q_B}$. Demand would expand along OB until the critical mass was reached. Then, for further expansion price should be fixed at $p_A(-\epsilon)$, and demand would expand along BA.

Unfortunately, the equal-usage assumption may not be very realistic. If not, usage-proportional pricing would exclude individuals with a lot of low value usage, and would admit (for a low price) individuals with a small amount of high value usage.

This might not be a bad idea, even in the long run. In general, usage-proportional pricing would be appropriate if the network were being used at capacity, and costs were closely related to total usage, not necessarily to the number of subscribers. A fixed price per subscriber would be appropriate if the network were not being used at capacity, and much equipment had to be committed for each subscriber.

The latter is perhaps more typical of a new communications service. Thus, usage-proportional pricing would probably result in some inefficiencies and misallocation of resources. Whether these inefficiencies are substantial or trivial is an empirical question. In any event, they should be compared to the inefficiencies of other start-up programs; e.g., the direct approach discussed in the previous subsection.

For a mature service the price should probably have a higher fixed component and a lower usage component. However, the externalities in consumption still have to be taken into account.[14]

A program with a low introductory price relies more on market processes than does the direct approach discussed in the previous subsection and does not depend so critically on the managers' judgment. In particular, the managers are not required to determine which individuals have the highest w_i. The individuals select themselves by choosing to subscribe to the service at the offered price.

☐ **Community of interest groups.** Community of interest groups may greatly reduce the practical difficulty of starting up the service. Maximum equilibrium demand may be achieved even if the initial user set is small—so long as that set exceeds the critical masses for some community of interest groups.

At the same time, community of interest groups place a greater burden on whatever procedure is used to select the initial users. If the initial user set is selected by managers, they must know what the community of interest groups are and decide how many individuals to select from each. In some circumstances, it would be optimal to select everyone from the same group; in other cases a more even spread would be optimal. In any event, the managers must make this choice, and the success of the program may be greatly influenced by how well they choose.

Community of interest groups also place a greater burden on the market process for programs involving a low introductory price. The expansion path involves price and $g_1, \ldots g_k$ (instead of q as in Figure 4). Efficiency requires that k different linear combinations of the g_i be concave downward. (The equivalent condition to γ's being concave downward in Figure 4.) But we have only one control variable, p. Clearly, it may be impossible to satisfy all k concavity conditions.

Thus, *any* program based on a (single) low introductory price may be inefficient in the sense that some individuals may join at the low introductory price but later drop out as the price rises. If this problem is serious, the seller may find it advantageous (and perhaps necessary) to use discriminatory pricing to assure that only "permanent" users join.

☐ **Few substantial contacts.** If an individual's demand depends on his principal contacts' being users, the start-up problem may be fundamentally different from that discussed above. It may be unnecessary for there to be hundreds or thousands of users before an individual demands the service. A user population of two or three may be self-sufficient—if they are the right two or three people.

---

[14] See Squire [3].

These small self-sufficient sets do not necessarily promote further growth, but they do allow a different kind of approach to starting up the service. The seller can begin by establishing small self-sufficient user sets. He can then gradually expand the size and number of these sets until the desired equilibrium user set is attained or until the service starts to grow by itself. The practicality of this method depends primarily on the size of the (minimum) self-sufficient user sets.

The smallest possible self-sufficient user sets consists of two mutual contacts for whom both v_{ij} and $v_{ji} \geqslant p$. If there are many such pairs i, j in the population, the service will start up, expanding beyond the minimum equilibrium set, with little or no help. Selling the service to i and j requires only getting the two together. And they may organize themselves and agree to subscribe to the service (contrary to the disequilibrium adjustment process assumed above[15]). After i and j both subscribe to the service, they may attract other individuals, and the service can grow further.

In fact, such growth from self-sufficient sets of two probably accounts in large part for the success in starting up telephone service. Indeed, telephone made a substantial penetration of the market while it was entirely a private-line service.

However, even if the service is viable, its incremental utility may be insufficient for very many people to demand the service to communicate with a single principal contact. If this is the case, the start-up problem is more difficult, and we must deal with larger self-sufficient user sets.

A self-sufficient set of three mutual contacts might also be able to organize itself and have all three members agree to subscribe together. However, this becomes progressively more difficult and unlikely as the size of the self-sufficient sets increases. It would be especially difficult if all the members of the set were not mutual contacts and no one knew all the other members of the set.

The seller might be able to gather data and determine self-sufficient user sets. He could then try to sell the service to everyone in such a set simultaneously. Naturally this is more difficult, the larger is the self-sufficient set. In fact, the difficulties of organizing even six to eight people and getting them all to agree to a joint purchasing decision may be far from trivial.

Nevertheless, the seller might do well to gather data on communications patterns and try to determine self-sufficient user sets. But it may be necessary to combine this with some other kind of start-up program. This could take the form of the direct approach or the low introductory price previously discussed. However, it is also possible to have a continual program in which each new subscriber is offered a low rate until the seller can connect a self-sufficient set of users that contains him. This kind of program would be effective if new users had high values of w_t but little community of interest with the current user set.

---

[15] P. 22.

References

1. Artle, R. and Averous, C. "The Telephone System as a Public Good: Static and Dynamic Aspects." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 89–100.
2. Marris, R. *The Economic Theory of "Managerial" Capitalism.* New York: Free Press, 1964.
3. Squire, L. "Some Aspects of Optimal Pricing for Telecommunications." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 2 (Autumn 1973), pp. 515–525.

Rate of return regulation and efficiency in production: an empirical test of the Averch-Johnson thesis

Robert M. Spann

Assistant Professor of Economics
Virginia Polytechnic Institute and State University

Eleven years ago, Harvey Averch and Leland Johnson demonstrated that firms subject to rate of return regulation have an incentive to use inputs in proportions that differ from the cost minimizing input levels. In the intervening years, numerous theoretical papers have appeared in the literature refining their original thesis, but, until recently there has been little or no empirical testing of this proposition. In this paper a test of the Averch-Johnson thesis using the trans-log production function is derived. The test is applied to data from regulated electric utilities and the Averch-Johnson thesis is confirmed.

1. Introduction

■ The pricing and output decisions of many utility firms are subject to review by state regulatory agencies.[1] The most common form of regulation is rate-of-return regulation in which the firm is constrained to pricing policies that produce a rate of return on assets equal to or less than some predetermined level.[2] Averch and Johnson, Wellisz,[3] and numerous others have argued that such regulation will cause the regulated firm to use inputs in proportions which differ from the cost-minimizing factor combinations.[4] Specifically, it will pay the regulated firm to be more capital-intensive at any given level of output than would be consistent with cost minimization.[5]

Robert M. Spann holds the B.S. degree in economics from North Carolina State University (1969) and the Ph.D. in economics and statistics from that institution (1973). His current research interests include the effects of public policy on industrial performance, the economics of energy markets, and the application of economic models to the regulatory process.

The author is indebted to Dale W. Jorgenson, Paul W. MacAvoy, Vittorio Bonomo, Thomas Hogarty, Warren Weber, and Ernst Berndt for helpful comments and criticisms on earlier drafts of this paper.

[1] While the nature and scope of regulation vary from state to state, every state has some form of regulatory bodies, and all states but two (Texas and Minnesota) regulate electric utilities.

[2] Most states only require that firms be constrained by regulators to earn a "fair" rate of return. The actual allowed rate of return is often the subject of lengthy proceedings in rate hearings.

[3] See [1] and [19], respectively.

[4] For a review of this literature, see Baumol and Klevorick [2].

[5] The Averch-Johnson thesis has other implications for the behavior

Despite its implications for policy, the Averch-Johnson hypothesis has been subject to few direct tests.[6] In what follows, a method of testing the Averch-Johnson thesis is proposed. The test is applied to data from regulated electric utilities and the Averch-Johnson thesis is confirmed.

2. Testing the Averch-Johnson thesis

■ The overcapitalization in regulated firms hypothesized by Averch and Johnson is a direct result of a model which starts with the premise that the regulated firm maximizes profits subject to an effective rate-of-return constraint.

The method used to test the Averch-Johnson thesis is to derive the restrictions on the factor share equations implied by the Averch-Johnson thesis. These restrictions can then be tested by using conventional statistical techniques in order to determine the empirical validity of the Averch-Johnson model. The restrictions used to test the Averch-Johnson thesis are a direct result of the equilibrium conditions for the profit-maximizing firm. These equilibrium conditions are reviewed briefly below. Then the actual method used to test the Averch-Johnson thesis is derived.

□ **Equilibrium conditions for the regulated monopoly.** The Averch-Johnson model assumes that the regulated firm produces a homogeneous output using two or more inputs. In the case of electricity there are three inputs: capital, labor, and fuel. The regulatory body constrains the firm in only one way: the firm is permitted to earn no more than a fixed percentage of its capital investment or a "fair rate of return" on its rate base. Finally, it is assumed that the allowed rate of return is less than the rate of return the firm would obtain if it were allowed to maximize profits, but is greater than the firm's cost of capital.

If we define

Q = the firm's output,
K = the amount of capital employed by the firm,
L = the amount of labor employed by the firm,
F = the amount of fuel used by the firm,
P = the price of the firm's output,
w = the wage rate,

---

of the regulated firm. In their original article, Averch and Johnson [1] argued not only that regulated firms would overcapitalize but that cross-subsidization may be profitable under regulation to the extent that it allowed the firm to expand its rate base. In a more recent paper Johnson [8] has argued that regulation may lead firms to buy rather than lease equipment. Scherer [13] has constructed a numerical example which implies that the "tighter" the regulation, the greater the social cost of producing electricity. A similar result is contained in Baumol and Klevorick [2]. Westfield [21] has argued that rate-of-return regulation may actually lead firms to overpay for capital equipment.

[6] Moore [11] attempted to test the Averch-Johnson thesis by comparing the input choices made by investor-owned and public utilities. Unfortunately, there is no reason to believe that municipal utilities will choose optimal input combinations. Courville [6] also presents a test of the Averch-Johnson thesis. His methodology differs significantly from the approach taken in this paper.

q = the price of fuel,
r = the cost of capital, and
s = the allowed rate of return,

then the regulatory constraints may be written as

$$sK \geqslant PQ - wL - qF. \tag{1}$$

The firm seeks to choose levels of K, F, and L to maximize profits

$$\pi = PQ - wL - rK - qF, \tag{2}$$

subject to the regulatory constraint given by equation (1).

If the regulatory constraint is assumed to hold as an equality, the firm's decision problem can be expressed mathematically by the Lagrangian

$$\begin{aligned} &\operatorname*{Max}_{K,L,F} R(K,L,F) - wL - rK - qF \\ &- \lambda[R(K,L,F) - wL - qF - sK], \end{aligned} \tag{3}$$

where $R(K, L, F)$ is the regulated firm's revenue function.[7]

First-order conditions are determined by differentiating equation (3) with respect to K, L, F, and λ. The necessary conditions for a maximum are:

$$R_K - r = \frac{\lambda}{1-\lambda}(r - s) \tag{4a}$$

$$R_F = q \tag{4b}$$

$$R_L = w \tag{4c}$$

$$R(K,L,F) = sK + wL + qF, \tag{4d}$$

where the subscripts K, L, and F indicate partial differentiation with respect to capital, labor, and fuel, respectively.[8]

Since Baumol and Klevorick have shown that $0 < \lambda < 1$, equation (4a) implies that the regulated firm is too capital intensive in the sense that, in equilibrium, the marginal revenue product of capital is less than the cost of capital.[9]

In the next section, a specific production function (the transcendental-logarithmic production function) is assumed for electric utilities. It is shown that this production function, in conjunction with equations (4a) thru (4d), implies a specific set of constraints on the factor share equations. The estimating equations involve λ, the Lagrangian multiplier for the rate of return constraint, and thus a value for λ can be estimated. The A-J thesis implies that

---

[7] Total revenue is $P \cdot Q$, which can be written as $\phi(Q) \cdot f(K, L, F) = \phi[f(K, L, F)] \cdot f(K, L, F) = R(K, L, F)$, where $f(K, L, F)$ is the firm's production function and $\phi(Q)$ is the inverse demand function.

[8] For a discussion of the second-order conditions, see Baumol and Klevorick [2].

[9] It should be noted that the A-J thesis does not imply that regulated firms engage in gold-plating or explicit rate base padding. The production function is assumed to be the same whether the firm is regulated or not. The overcapitalization implied by this model consists solely of inputs being used in proportions that differ from the cost minimizing input combinations. There is no wastage of inputs in a physical sense, only inefficient usage of those inputs in an economic sense. For a discussion of the comparative statics of the A-J model, see Baumol and Klevorick [2].

λ is between zero and one, and this hypothesis can be tested using the framework developed below. If λ is not significantly different from zero, this implies that the rate-of-return constraint does not enter the firm's objective function and the A-J thesis is invalid. In addition, the model implies certain restrictions on the estimating equations which can be used to test the A-J thesis.

☐ **A test of the A-J thesis using the transcendental-logarithmic production function.** In order to test the Averch-Johnson thesis, I use the transcendental-logarithmic (trans-log) production function derived by Christensen, Jorgenson, and Lau.[10] This production function is a very general production function of which many commonly used production functions (the Cobb-Douglas, CES, etc.) may be viewed as special cases. Essentially it is a Taylor series expansion of any arbitrary production function.[11]

There are three inputs in the production of electricity: capital, labor, and fuel. The three-input trans-log production function may be written as[12]

$$\begin{aligned}\log Q = \log \alpha_0 &+ \beta_1 \log K + \beta_2 \log L + \beta_3 \log F + \beta_4 (\log K)^2 \\ &+ \beta_5 (\log L)^2 + \beta_6 (\log F)^2 + \beta_7 \log K \log F \qquad (5) \\ &+ \beta_8 \log K \log L + \beta_9 \log F \log L.\end{aligned}$$

By taking the logarithmic derivative of equation (5) with respect to capital, one obtains

$$\frac{\partial \log Q}{\partial \log K} = \beta_1 + 2\beta_4 \log K + \beta_7 \log F + \beta_8 \log L. \qquad (6)$$

By definition

$$\frac{\partial \log Q}{\partial \log K} = f_K \frac{K}{Q},$$

where f_K is the marginal product of capital. If we assume the demand curve facing the regulated monopolist is a constant elasticity demand function, equation (4a) may be rewritten as

$$P\left(1 + \frac{1}{\eta}\right) f_K = r + \frac{\lambda(r - s)}{(1 - \lambda)}, \qquad (7)$$

where η is the elasticity of demand. Substituting equation (7) into equation (6) and rearranging terms, one obtains an estimable equation of the form

$$\mu_K = b_1 + b_2 \log K + b_3 \log F + b_4 \log L + \lambda Z, \qquad (8)$$

where

$\mu_K = \dfrac{rK}{PQ}$ or the required payments to capital (the rental rate of capital times capital used) as a fraction of total revenue and

---

[10] In [5].

[11] For other applications of the trans-log production function, see Berndt and Christensen [3] or Binswanger [4].

[12] The usual discussion of the A-J thesis assumes a two-input production function. Addition of a third input does not alter the implications of the A-J model.

$Z = \frac{sK}{PQ}$ or the allowed payments to capital (the allowed rate of return times capital used) as a fraction of total revenue.[13]

One could test the A-J model by estimating equation (8) and by testing the null hypothesis, $\lambda = 0$, or that the regulatory constraint does not enter the firm's objective function. Such a procedure would not use all the information the data contain.[14] The trans-log production function and the A-J thesis also imply a specific equation for the shares of other factors and specific constraints on those equations.

If one takes the logarithmic derivative of the trans-log production function with respect to F, fuel input, the result is

$$\frac{\partial \log Q}{\partial \log F} = \beta_3 + 2\beta_6 \log F + \beta_7 \log K + \beta_9 \log L = f_F \frac{F}{Q}, \quad (9)$$

where f_F is the marginal product of fuel. Substituting into the first-order condition for fuel, one obtains[15]

---

[13] If equation (7) is substituted into equation (6), one has

$$\frac{K\left(r + \frac{\lambda r}{1-\lambda} - \frac{s\lambda}{1-\lambda}\right)}{PQ\left(1 + \frac{1}{\eta}\right)} = \beta_1 + 2\beta_4 \log K + \beta_7 \log F + \beta_8 \log L.$$

Rearranging terms,

$$\frac{rK}{PQ} = \frac{\left(1+\frac{1}{\eta}\right)\beta_1}{1+\frac{\lambda}{1-\lambda}} + \frac{2\left(1+\frac{1}{\eta}\right)\beta_4 \log K}{1+\frac{\lambda}{1-\lambda}} + \frac{\left(1+\frac{1}{\eta}\right)\beta_7 \log F}{1+\frac{\lambda}{1-\lambda}}$$

$$+ \frac{\left(1+\frac{1}{\eta}\right)\beta_8 \log L}{1+\frac{\lambda}{1-\lambda}} + \frac{\lambda}{(1-\lambda)}\left(\frac{1}{1+\frac{\lambda}{1-\lambda}}\right)\frac{sK}{PQ}.$$

The coefficient of $\frac{sK}{PQ}$ reduces to λ, and thus equation (8) follows with

$$b_1 = \frac{\left(1+\frac{1}{\eta}\right)\beta_1}{1+\frac{\lambda}{1-\lambda}}, \quad b_2 = \frac{2\left(1+\frac{1}{\eta}\right)\beta_4}{1+\frac{\lambda}{1-\lambda}},$$

$$b_3 = \frac{\left(1+\frac{1}{\eta}\right)\beta_7}{1+\frac{\lambda}{1-\lambda}} \text{ and } b_4 = \frac{\left(1+\frac{1}{\eta}\right)\beta_8}{1+\frac{\lambda}{1-\lambda}}.$$

[14] The information contained in equation (8) is not useless, however. Versions of this equation were estimated and they are discussed below.

[15] Note that

$$\mu_F = b_3 + b_6 \log K + b_7 \log F + b_8 \log L, \qquad (10)$$

where $\mu_F = \dfrac{qF}{PQ}$ of payments to fuel as a fraction of total revenue.

Equations (8) and (10) form a system of two equations which can be used to estimate λ. All the parameters of these two equations are not functionally independent. If b_6 is divided by b_3, one obtains

$$\frac{b_6}{b_3} = \frac{\left(1 + \dfrac{1}{\eta}\right)\beta_7}{\left(1 + \dfrac{1}{\eta}\right)\beta_7 \Big/ 1 + \dfrac{\lambda}{1 - \lambda}} = 1 + \frac{\lambda}{1 - \lambda}$$

Upon rearrangement this becomes

$$(1 - \lambda) b_6 = b_3. \qquad (11)$$

Equation (11) is a constraint on the two equation system [equations (8) and (9)] for factor shares that is implied by profit maximization subject to the regulatory constraint. Thus we should be able to use this constraint to derive a test of the Averch-Johnson thesis.

☐ **Testing the A-J constraints.** The Averch-Johnson thesis states that firms maximize profits given the rate-of-return constraint. The Averch-Johnson thesis can fail to apply to regulated firms if either or both of two conditions hold:

(1) the regulatory constraint does not enter the firm's objective function, or
(2) regulated firms do not maximize profits.

Condition (1) implies the following testable hypothesis

$$H_1 : \lambda = 0.$$

Condition (2) implies

$$H_2 : (1 - \lambda) b_6 \neq b_3.$$

Unfortunately, both H_1 and H_2 cannot be tested simultaneously. No matter how the joint hypothesis H_1 and H_2 (or its complement) is expressed, the alternative hypothesis will always contain a single point alternative. This does not mean that the constraint implied by equation (11) cannot be used to test the A-J thesis. One can make tests of either proposition 1 or proposition 2, given that the other (or its complement) holds.

Suppose one wanted to test the hypothesis that the regulatory constraint enters the firm's objective function, given that firms maximize profits. This hypothesis would be tested by estimating equations (8) and (10) jointly, subject to the constraint,

---

$$b_5 = \left(1 + \frac{1}{\eta}\right)\beta_3,\ b_6 = \left(1 + \frac{1}{\eta}\right)\beta_7,$$

$$b_7 = 2\left(1 + \frac{1}{\eta}\right)\beta_6 \text{ and } b_8 = \left(1 + \frac{1}{\eta}\right)\beta_9.$$

$$(1-\lambda)\, b_6 = b_3,$$

and by testing the hypothesis $\lambda = 0$. In other words, we test the hypothesis $\lambda = 0$ (the regulatory constraint does not enter the objective function) within the constraints on other parameters implied by profit maximization.

Alternatively, one could test the hypothesis that regulated firms maximize profits, given that the regulatory constraint enters the objective function. This hypothesis would be tested by estimating equations (8) and (10) jointly assuming $\lambda \neq 0$ (i.e., Z is included as a variable) and testing the nonlinear restriction,

$$(1-\lambda) b_6 = b_3.$$

If this restriction is not rejected, we accept the hypothesis that firms maximize profits, given an effective regulatory constraint.

In what follows, both approaches to testing the A-J model are used. In almost all cases, the existence of the A-J effect cannot be rejected.[16]

There is a problem with the tests discussed here which does need to be mentioned. The term λ is treated as a parameter when it is really a variable within the model. The equations estimated are a subset of a more detailed simultaneous equation model which would include not only the share equations reported here, but the output demand equation and the production function as well.[17] Within such a model λ could be treated as a variable, taking a different value for each observation in the sample. Unfortunately, λ is not directly observable. Thus, in the estimations that follow, the estimates of λ should be viewed as estimates of $\bar{\lambda}$, or the average of λ for the sample.[18]

---

[16] These results must be interpreted with care. The two alternative tests discussed here are both applied to the same set of data. The fact that both tests imply acceptance of the A-J thesis cannot be interpreted as implying two independent confirmations of the A-J thesis. The tests are clearly not independent and thus the test statistics used to test the A-J thesis do not have the properties usually attributed to those statistics. In addition, the same data are used in this section as were used in the previous section. Thus the test statistics presented in the following section are not independent of the results of the previous section. While this does reduce the robustness of the tests used and while it would be preferable to have several sets of independent data to perform these tests, it is worthwhile to determine whether different approaches to the same problem lead to similar conclusions. The fact that different approaches to testing the A-J thesis all imply the same conclusion is worth something.

[17] The identity $Z + \mu_L + \mu_F \equiv 1$ (where μ_L is payments to labor as a fraction of total revenue) derived from the regulatory constraint implies that any attempt to estimate all three share equations simultaneously would lead to a singular error matrix. One would estimate the production function in addition, if he felt that the error term in the share equations included not only stochastic elements in the production function, but also random derivations from strict profit maximization due to planning errors, etc.

[18] One alternative to assuming that λ does not vary greatly, or significantly, across firms would be to estimate the parameters of the production function from a simultaneous estimation of the production function and the labor and fuel equations. Given these estimated parameters, one could use equation (8) to estimate a $\hat{\lambda}$ and a variance of $\hat{\lambda}$ for each firm. In some preliminary regressions, this was done and $\hat{\lambda}$ did not vary much within the data used here.

In the next section the data used to test the Averch-Johnson thesis are described. Then the tests of the model are presented.

3. Data requirements

■ At first glance, the data necessary to test the Averch-Johnson thesis appear to be so abundant as to make one wonder why no empirical papers on the Averch-Johnson thesis have appeared in the literature. Telephone and electric power companies are regulated in almost every state and there are at least two firms in each of these industries in every state.[19] Therefore, one has at least 100 observations available for econometric tests of the A-J thesis. Furthermore, federal agencies publish volumes of annual data on these firms.[20]

Unfortunately, several complications, primarily technical, prevent use of all these data. With regard to the telephone industry, it is not clear exactly what the industry's output is, nor are there well-defined methods of handling the problem of long distance vs. local calls, party lines vs. private lines.[21]

Electrical power companies actually produce three distinct outputs, each with different production functions: the generation of electric power, the transmission of electric power, and the distribution of electric power. Different firms are involved in these activities in varying degrees.

The rate of adjustment to input price changes and demand shifts in the electric power industry may be fairly slow due to the long lead times involved in plant construction and the capital intensity of the industry. Thus, any data used must come from a relatively "stable" time period, in order to insure that our observations represent points of long-run equilibrium.

Two sets of data on electric power companies satisfied the restrictions implied by the above discussions. The first data set consists of observations on the first year of operation for all the new steam electric plants built by regulated firms between 1959 and 1963.[22] This period represented a time of fairly stable input prices and is prior to unexpected demand increases, increasing coal costs, and lengthier regulatory proceedings of the late 1960s and early 1970s. This sample consists of 35 plants.

The second set of data consists of a select sample of large electric companies in 1963. The sample was restricted to firms for which non-hydro-electric plants represented at least 90 percent of generating capacity, at least 90 percent of kilowatt-hour sales were generated by the firm instead of purchased from others, and, no more than 10 percent of kilowatt-hour generation was resold to other power companies.[23] These criteria were satisfied by 24 regulated companies.

---

[19] The only states which do not regulate telephone and electric power firms at the state level are Texas and Minnesota.

[20] In fact, the Federal Power Commission's *Statistics of Privately Owned Electric Utilities in the United States* lists 163 class A and B (firms with sales of two million kilowatt-hours or more) electric power firms.

[21] The primary problem of telephone output is that the use of a phone involves the privilege both to place and receive calls. Thus, the probability of being able to complete a call is part of the output of the phone company.

[22] Technically, these plants should be called fossil plants, since steam electric plants may be either nuclear or fossil fueled. However, during the time period under consideration, virtually all steam plants were fossil fueled so the terms fossil plant and steam plant will be used interchangeably.

☐ **Definition of variables.** In order to apply our test of the A-J thesis, we need observations on total revenue, the rate of return constraint, the rental rate on capital, and the three factor inputs, labor, capital, and fuel. The definition of each variable is discussed in turn.

☐ **The rental rate on capital and the constrained rate of return.** In order to estimate the rental rate on capital (r) and the allowed rate of return (s), one needs estimates of allowed rates of return, the cost of capital, the corporate tax rate, the rate of depreciation, and the regulated firm's capital structure.[24] For the purposes of this paper, the cost of capital was assumed to be the same for all firms and equal to the average cost of capital to electric utility firms derived by Litzenberger and Rao of 5.6 percent.[25] The estimate of the corporate tax rate is 0.52, the estimate of the depreciation rate is 0.04 (provided by industry sources) and the estimate of the debt-equity ratio is 0.52, the average portion of debt in electric

---

[23] The purpose of these restrictions is fairly obvious. The production function is not the same for steam generation and hydroelectric generation. A higher fraction of sales for resale implies a greater investment in generation and transmission relative to distribution. A higher fraction of purchased power leads to the opposite result Thus, in order to insure homogeneity of the parameters of the production function for the firm, the sample must be restricted to firms with similar values for hydro vs. steam capacity, purchased vs. generated power, and resale vs. retail sales. In addition, some firms which satisfied these constraints were not included due to incomplete data.

[24] Although taxes are not included in the previous analysis of the A-J model, it can be shown that inclusion of a corporate income tax has the effect of changing the rental rate on capital to

$$r = c\left[\frac{i}{(1-t)} + \delta - \frac{pti}{(1-t)}\right]$$

where

c = the price of capital goods
i = the cost of funds
δ = the depreciation rate on capital
t = the income tax rate
p = the fraction of capital costs deductible from taxable income.

The constrained rate of return, in the presence of an income tax, becomes

$$s = c\left[\frac{i'}{(1-t)} + \delta - \frac{pti}{(1-t)}\right],$$

where i' is the allowed rate of return.

[25] In [10]. The assumption of a constant cost of capital to all electric utilities is not quite correct. However, debt-equity ratios varied little among firms (the standard deviation of the equity portion of firm's capital structure was about 1/8 the mean in the firm data). Beta factors vary little and debt costs vary less among utilities than among industrial firms. Thus, the assumption of a constant cost of capital is probably reasonably accurate. Some attempts were made to estimate separate costs of capital for each firm. This did not alter the results of the tests reported below.

utilities' capital structure. The corresponding estimate of s is the average rate of return for each firm for the period 1961–1963 (1961 was the earliest year the FPC provided these data) as computed by the FPC.[26]

☐ **The capital stock.** Two measures of the capital stock are used. The first is simply the megawatt capacity of the plant or firm (denoted by K_1). The second (denoted by K_2) is the book value of total assets (in the case of the firm) or the book value of plant, equipment, and land (in the case of the plant data). Which measure is more appropriate depends on whether differing costs per megawatt capacity reflect differing construction costs as opposed to quality differences. Which measure is more appropriate cannot be determined *a priori*.[27]

☐ **Labor.** The firm's or plant's labor usage is defined to be the total number of employees.

☐ **Fuel.** Fuel is total BTUs of fuel consumed by the firm (or plant).

☐ **Total revenue.** Total revenue is simply total revenue as reported by the firm in the case of the firm data. For plant data, total revenue is estimated as the sum of s times cost of plant, fuel expense, and costs other than fuel.

☐ **The wage rate.** The wage rate, w, is payments to labor divided by total employees.

☐ **The price of fuel.** The price of fuel, q, is total fuel costs divided by millions of BTUs of fuel burned.

☐ **Factor payments as a fraction of total revenue.** Payments μ_K and μ_F are $\frac{rK}{PQ}$ and $\frac{qF}{PQ}$, respectively.

4. Empirical estimates of the Averch-Johnson effect

■ In Table 1, ordinary least squares estimates of equations (8) and (10) subject to the constraint given by equation (11) are listed. This is the basic estimating equation which is used for the tests of the Averch-Johnson thesis. In those equations, the subscripts 1 and 2 on capital refer to whether megawatt capacity or total assets were used to estimate the size of the capital stock. The method of estimation was to minimize the sum of errors squared around both equations by using a nonlinear search procedure for λ in the interval $(0, 1)$.[28] No standard error is shown for λ (the coefficient of Z) because of the nonlinear procedure used.

---

[26] The assumption implicit in using a three-year average of earned rates of return as a proxy for allowed rates of return is that regulated firms were subject to periodic, although not perfect or instantaneous, review during this period. If this is the case, the average rate of return over a period of years should be close to the allowed rate of return.

In the data actually used, allowed rates of return appeared to be about 25 percent above required rates of return (or the cost of capital). This divergence between s and r may seem high by current standards, but one must remember that the early 1960s was a period in which electric utility rates were either falling, or rising much less than they are today. In addition, the higher returns earned by the owners of electric utility stocks in the early 1960s (relative to current returns) might indicate a downward trend in the divergence between s and r.

[27] When K_1 is used as the measure of the capital stock, the cost of a unit of capital equipment (c) is the ratio of K_2 to K_1.

TABLE 1

RESTRICTED REGRESSIONS USED TO TEST THE A-J THESIS*

	FIRM DATA
(1)	$s_{K_e} = 0.3796 \underset{(0.02657)}{-0.03093} \log K_1 + \underset{(0.01965)}{0.0348} \log L - \underset{(0.01145)}{0.00865} \log F + 0.6848Z$
	$s_F = -0.00682 - \underset{(0.03632)}{0.02744} \log K_1 + \underset{(0.01975)}{0.00389} \log L + \underset{(0.03489)}{0.02870} \log F$
	$R^2=0.6506$ SSE $= 0.04789$
(2)	$s_K = 0.11233 - \underset{(0.01101)}{0.01493} \log F + \underset{(0.02228)}{0.01028} \log L + \underset{(0.029503)}{0.007623} \log K_2 + 0.66174Z_e$
	$s_F = 0.35298 - \underset{(0.032556)}{0.044164} \log K_2 + \underset{(0.022195)}{0.014917} \log L + \underset{(0.028535)}{0.03067} \log F$
	$R^2=0.5664$ SSE $= 0.049371$
	PLANT DATA
(3)	$s_K = 0.95150 - \underset{(0.02914)}{0.010259} \log L + \underset{(0.028520)}{0.05670} \log K_1 - 0.06614 \log F + 0.56504Z$
	$s_F = -1.2388 - 0.15208 \log K_1 + \underset{(0.05853)}{0.15241} \log F - \underset{(0.044143)}{0.01781} \log L$
	$R^2 = 0.7849$ SSE $= 0.2180$
(4)	$s_K = 0.69630 - \underset{(0.02884)}{0.04258} \log L + \underset{(0.02637)}{0.09481} \log K_2 - \underset{(0.19976)}{0.08245} \log F + 0.50094Z$
	$s_F = 0.23859 - \underset{(0.03995)}{0.1649} \log K_2 + \underset{(0.03858)}{0.13371} \log F + \underset{(0.04467)}{0.03412} \log L$
	$R^2 = 0.75582$ SSE $= 0.18381$

*SSE = SUMS OF ERROR SQUARED. SSE AND R^2 REFER TO THE SUMS OF ERROR SQUARED ABOUT BOTH REGRESSION EQUATIONS STANDARD ERRORS ARE IN PARENTHESES

The plant regressions were weighted by the inverse of the standard errors of estimate of the unconstrained fuel and capital share equations, since these variances were significantly different.

Two hypotheses are tested using these data. The first is that, given that the symmetry constraint holds, $\lambda = 0$. This is tested by computing the χ^2 statistic:

$$\chi_1{}^2 = -T \log \frac{L(H_A)}{L(H_0)},$$

where

T = the number of observations,

$L(H_A)$ = the sum of errors squared under the alternative hypothesis, $\lambda \neq 0$, and

$L(H_0)$ = the sum of errors squared under the null hypothesis, $\lambda = 0$.

---

[28] An alternative method of estimation would be an instrumental variables approach since K, L, and F are endogenous to the firm. Observations on exogenous variables such as income, population, and prices of alternative fuels (such as gas for home heating) are not available on an electric utility service area basis, and thus there are not enough exogenous variables to identify equations (8) and (10).

Acceptance of this null hypothesis would indicate that the Lagrangian multiplier for the rate-of-return constraint was zero. This would imply that regulation is not effective and that the A-J effect is inoperative.

The χ^2 statistics for testing $\lambda = 0$ for each estimation are presented in Table 2. The column headed Data Set in Table 2 refers to which sample (firm or plant data) was used. As can be seen, λ is significantly different from zero at the 0.01 level in all cases. In addition, the estimated values of λ are within the range predicted by the A-J model.[29] Inspection of Table 1 indicates that all the estimated λ's are between zero and one. The range of estimated λ's is 0.5–0.7.[30]

TABLE 2

χ^2 STATISTICS FOR TESTING $\lambda=0$

DATA SET	MEASURE OF CAPITAL STOCK	ESTIMATED	χ^2 TO TEST $\lambda=0$
PLANT	K_1	0.5650	26.48*
PLANT	K_2	0.5009	25.77*
FIRM	K_1	0.6848	43.98*
FIRM	K_2	0.6617	34.43*

*SIGNIFICANT AT 0.01 LEVEL

In Table 3, the χ^2 statistics for testing the profit maximizing or symmetry constraint

$$(1 - \lambda)\, b_6 = b_3$$

are listed.[31] The estimations are conducted under the assumption $\lambda \neq 0$. As can be seen, the constraint is accepted in the plant data. In the firm data, the χ^2 statistics are below the critical level when K_1 is the measure of the capital stock but not when K_2 is used. This would seem to indicate confirmation of the A-J thesis in the plant data but raise some doubts concerning its validity in the firm data.

TABLE 3

χ^2 STATISTICS FOR TESTING PROFIT-MAXIMIZATION RESTRICTION GIVEN EFFECTIVE REGULATION

DATA SET	MEASURE OF CAPITAL STOCK	χ^2
PLANT	K_1	0.987
PLANT	K_2	1.414
FIRM	K_1	3.586
FIRM	K_2	9.877*

*SIGNIFICANT AT 0.01 LEVEL

The rejection of the symmetry constraint in the firm data may be due less to a failure of the Averch-Johnson thesis than to other problems with the firm data. As noted earlier, the production and sale of electricity is a three-stage process. This could introduce aggregation problems, in which case the symmetry constraint would fail for reasons independent of the Averch-Johnson thesis.[32] In addition, there may be errors in the measurement of the capital stock due to sale and lease-back arrangements. Such arrangements are not uncommon in this industry.[33] Finally, transmission and distribution plant decisions may be influenced by political and environmental constraints as well.

---

[29] It will be recalled that the A-J model requires $0 < \lambda < 1$.

[30] A similar range of λ's was observed when only the capital equation was estimated.

[31] The null hypothesis is $(1 - \lambda)\, b_6 = b_3$ with the alternative being $(1 - \lambda)\, b_6 \neq b_3$.

[32] Since the output is produced via a multistage process, a production function expressed as output as a function of the aggregate levels of the inputs may not be well-defined at the firm level.

[33] Under a sale and lease-back arrangement, the firm has use of the same amount of capital, yet its recorded assets decrease and current expenses increase. An additional problem with the firm data is that the pro-

Given the caveats stated above, the evidence presented so far appears to be in favor of the Averch-Johnson thesis. The estimates of λ are significantly different from zero, which indicates that the regulatory constraint does enter the firm's objective function. In addition, the symmetry constraint holds in the plant data. Its failure in some of the firm data may be due to reasons independent of the validity of the Averch-Johnson thesis.

5. Summary and conclusions

■ A now classic hypothesis in the economics of regulated industries is that firms subject to rate-of-return regulation overcapitalize. There has been much theoretical discussion of this hypothesis, but little empirical testing of its merits. In this paper the Averch-Johnson thesis was expressed as a set of nonlinear restrictions on the factor share equations. These restrictions were tested by using plant and firm data from regulated electric utilities. The Averch-Johnson thesis was accepted in almost all cases. Where it was not accepted, the rejection of the Averch-Johnson thesis may be due to problems independent of the thesis itself.

The results of this study indicate that the evaluation of regulation may involve more than an examination of whether or not, or by how much, the existence of regulation alters the prices and outputs of the regulated firm. In evaluating the performance of regulation, one must include in his cost-benefit calculations the costs of input distortion when rate-of-return regulation is effective.[34]

Any benefits of electric utility regulation in terms of lower prices or some other measure of improved performance are not achieved at zero costs.[35] There are no free lunches. Imposing constraints on a firm leads that firm to alter its behavior in such a manner that profits are maximized subject to the regulatory constraint. If the profits of the regulated firm are constrained to be functionally related to its rate base, it pays the regulated firm to overcapitalize.

In addition, this study tests the hypothesis that regulatory bodies have some effect on the firm's decisions without assuming that regulation is equally effective in all states. By their use of a simple zero-one dummy variable for the absence or presence of regulation, Stigler and Friedland[36] assume that regulation is equally effective in all states. The present study, by varying the allowed

---

duction function at the firm level depends on population density. The more densely populated a utility's service area, the lower are transmission and distribution plant requirements. To the extent the omission of this variable leads to a specification bias, the tests presented here may be biased. Thus, the firm data should be considered less reliable than the plant data.

[34] This point has been made by Westfield [20] also. However, Westfield's analysis appears to go too far in the opposite direction. He appears to imply that estimates based on price changes and consumers' surplus are of little or no value. Our results only imply that in the special case of rate-of-return regulation, one must go beyond the data on prices and outputs.

[35] Not all regulation lowers price. For an example of a case in which the opposite is true, see Spann and Erickson [15].

[36] In [16].

rate of return, allows the tightness of regulation to vary from state to state.[37]

As Baumol and Klevorick[38] state, the Averch-Johnson thesis "provides a simple and comprehensive depiction of the regulatory process." That model is a powerful one and does yield implications which are testable and do appear to hold. It may be that the Averch-Johnson effects are not the most disquieting of the inefficiencies in the regulated sector.[39] This is beside the point here. The point is that economic analysis, even if it is in the form of relatively simple models, can explain behavior in the regulated sector. With the apparent trend for economists to return to close examination of the regulatory process, economic analysis has gained impetus in both the formulation and the application of regulation.[40]

References

1. Averch, H. and Johnson, L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. Baumol, W. and Klevorick, A. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
3. Berndt, E. and Christensen, L. "The Translog Function and the Substitution of Equipment, Structures, and Labor in U. S. Manufacturing." *Journal of Econometrics* (March 1973).
4. Binswanger, H. P. "The Measurement of Biased Efficiency Gains in U. S. and Japanese Agriculture to Test the Induced Innovation Hypothesis." Unpublished Ph.D. dissertation, North Carolina State University.
5. Christensen, L., Jorgenson, D., and Lau, L. "Conjugate Duality and the Transcendental Logarithmic Production Function." Unpublished paper presented at the Second World Congress of the Econometric Society, Cambridge, England, September 1970.
6. Courville, L. "Regulation and Efficiency in the Electric Utility Industry." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 1 (Spring 1974), pp. 53–74.
7. Donahue, C. "Lawyers, Economists, and the Regulated Industries: Thoughts on Professional Roles Inspired by Some Recent Economic Literature." *Michigan Law Review*, Vol. 70 (November 1971), pp. 195–220.
8. Johnson, L. L. "Behavior of the Firm under Regulatory Constraint."

---

[37] It is tempting to argue that this study offers an explanation of Stigler's finding that regulation of electric utilities had no effect on the price of electricity during the early years of regulation. Regulation could have had the effect of leaving price and output unchanged but raising costs by forcing firms to choose inefficient input combinations. However, this implies that profits of regulated firms should have fallen. Stigler found that profits were not affected by regulation. But on the other hand, acceptance of the Averch-Johnson thesis cannot be taken as a direct refutation of Stigler and Friedland [16]. The principal effect of regulation could be on income redistribution through cross-subsidization of the utilities customers, but it would not affect the level of average prices. For evidence that regulation alters the price structure of electric utilities, see Spann [14].

[38] In [2].

[39] For more on this point, see Baumol and Klevorick [2], p. 189.

[40] The fact that an increased usage of economists in the regulatory process is required has been recognized by lawyers as well. For example, see Donahue [7].

Paper presented at the Meetings of the American Economic Association, Toronto, 1972.

9. Jorgenson, D. "The Theory of Investment Behavior," in Robert Ferber, ed., *Determinants of Investment Behavior,* New York: National Bureau of Economic Research, 1967.
10. Litzenberger, R. and Rao, C. "Estimates of the Marginal Rate of Time Preference and Average Risk Aversion of Investors in Electric Utility Shares: 1960–66." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 265–277.
11. Moore, T. "The Effectiveness of Regulation of Electric Utility Prices." *Southern Economic Journal,* Vol. 36, No. 4 (April 1970), pp. 365–375.
12. Myers, S. "The Application of Finance Theory to Public Utility Rate Cases." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 58–97.
13. Scherer, F. M. *Industrial Market Structure and Economic Performance.* Chicago: Rand McNally and Company, 1971.
14. Spann, R. M. "Taxation, Regulation, Electric Utility Pricing and Efficiency in the Allocation of Energy Resources." Unpublished manuscript, Virginia Polytechnic Institute and State University.
15. ——— and Erickson, E. W. "The Economics of Railroading: The Beginning of Cartelization and Regulation." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 227–244.
16. Stigler, G. and Friedland, C. "What Can Regulators Regulate: The Case of Electricity." *Journal of Law and Economics,* Vol. 5, No 2 (October 1962).
17. U. S. Federal Power Commission. *Statistics of Privately Owned Electric Utilities in the United States.* Washington, D. C., published annually.
18. ———. *Steam-Electric Plant Construction Cost and Annual Production Expense.* Washington, D. C., published annually.
19. Wellisz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy,* Vol. 71, No. 1 (February 1963), pp. 30–43.
20. Westfield, F. "Methodology of Evaluation of Economic Regulation." *The American Economic Review,* Vol. 61, No. 2 (May 1971), pp. 211–217.
21. ———. "Regulation and Conspiracy." *The American Economic Review,* Vol. 55, No. 3 (June 1965), pp. 424–443.

Regulation and efficiency in the electric utility industry

Léon Courville

Assistant Professor
Graduate School of Management
The University of Rochester

Rate-of-return regulation is a distinct feature of the public utility environment. More than ten years ago, it was argued that this form of regulation might induce utility companies to operate inefficiently. This, the "Averch-Johnson effect," in spite of being the subject of widespread discussion, has not been assessed through formal quantitative analysis. This paper attempts such an analysis and presents evidence on the existence of the "A-J effect" by investigating a cross section of electric power plants. The Averch-Johnson proposition is strongly confirmed. Estimates of the magnitude of the regulation-induced inefficiency are also provided.

1. Introduction and summary

■ More than ten years ago Averch and Johnson[1] developed the proposition that the regulated monopolist will not be efficient in choosing its inputs. Rate-of-return regulation, by limiting the size of accounting profits per unit of capital to some value, known as the allowed rate of return, favors the use by the regulated monopolist of capital inputs, since it can increase its profits only if it increases the amount of capital. This proposition received a considerable amount of attention over the past decade. Much of the discussion of the Averch and Johnson proposition has been theoretical in nature. Yet, in spite of this, no formal quantitative evaluation of either the existence or importance of the A-J effect was attempted.[2] The absence of any empirical investigation is even

---

Léon Courville received the Lic. Sc. Com. from l'Ecole des Hautes Etudes Commerciales in 1967 and the M. A. in 1972 from the Graduate School of Industrial Administration, Carnegie-Mellon University, where he is currently a doctoral candidate. Professor Courville's current research is in regulation and applied microeconomics.

This paper is part of the author's doctoral dissertation to be presented to the faculty of the Graduate School of Industrial Administration, Carnegie-Mellon University. It is a revised version of a paper delivered at the Econometric Society Meetings, December 1972. This research was supported in part by the Center for Research in Government Policy and Business and by the Center for Naval Analysis Grant N00014-68-0091 to the University of Rochester. A Fellowship from l'Ecole des Hautes Etudes Commerciales, Montreal, also provided support to the author when this study began. The author wishes to thank Robert E. Lucas, Jr., Timothy McGuire, and Edward C. Prescott for their encouragement and good advice. He is also grateful to Kenneth Gaver and John Long for their helpful discussions and to Michael C. Jensen, Walter Oi, H. Martin Weingartner, and especially Paul W. MacAvoy for helpful comments on an earlier version. They are not responsible, of course, for any remaining errors.

[1] See Averch and Johnson [1].

[2] At least until very recently. See Spann [16].

more surprising, given the frequent appeal by some writers to the necessity of such investigation, given the casual empiricism to which it was subjected, and given its denial by public utility officials.

This paper is concerned exclusively with the quantification of the "A-J effect" in firms subject to rate-of-return regulation. Broader issues in regulation theory like the long-run effectiveness of a particular regulatory scheme, the relative merits of alternative regulatory schemes, and the social benefits of any regulatory control will not be discussed. The aim and underlying motivation of the study are to examine the Averch-Johnson proposition and to relate this hypothesis to the regulatory context it assumes so as to verify it empirically and to quantify its importance.

The material is organized as follows. In the next section, the Averch and Johnson model is briefly presented. The effect of a rate-of-return constraint on the monopolist's decision process is to render optimal, from the monopolist's point of view, a form of behavior which is inefficient. To be more precise, the monopolist will operate in such a way as to make the ratio of the marginal productivity of any capital input to the marginal productivity of any noncapital input smaller than the ratio of their respective prices. This proposition will be evaluated in the form just stated by investigating a cross section of firms in the electric utility industry.

The choice of electric utilities is taken up in the third section of the paper. A brief overview of the scope and nature of the regulatory environment of electric utilities is offered and it is argued that rate of return regulation is and has been the distinct feature of the regulatory process. Even though it is recognized that the workings of regulatory control are not adequately represented by the traditional Averch-Johnson analysis, mention is made of the modifications necessary to preserve the basic qualitative conclusion of that analysis. A brief description of a typical electric utility company leads to the fourth section of the paper where the approach for testing and evaluating the Averch-Johnson proposition is outlined. An electric company is basically engaged in three activities: the generation of electricity, its transmission, and its distribution. It is argued that in order to infer the existence of input distortion, one need only consider the generation of electricity. Pitfalls that could occur by looking at the firm as a whole are also mentioned. Alternative procedures for investigating the proposition stated above are briefly examined, and this discussion gives strong support for the approach taken in this paper: namely, to examine the relation between marginal productivities of the inputs to electricity generation and their respective prices.

To do this one must have an estimate of the production function in electricity generation. The concept of an *ex ante* production function, its relevance to the methodology, and the estimation problems it presents are then discussed. Specific industry characteristics like the measure of capital and the peak-load phenomenon are also discussed.

The "A-J proposition" is tested in the last section of the paper. The tests presented confirm the existence of input inefficiency (even when the methodology tends to bias the test against finding inefficiency). Finally, percentage deviations from the minimum cost combination are computed. The average deviation for the plants in the sample is estimated at 11.95 percent. Cost figures on the generation of electricity from the National Power Survey,[3] show that efficient electricity production would have saved 436.5 million dollars in 1962.

2. The Averch-Johnson proposition

■ The Averch-Johnson analysis provides a simple but comprehensive picture of the behavior of firms under regulation. Since this model has been widely discussed, it is not our intention to review all its aspects[4]; however, it is appropriate to consider its empirical proposition on input combination. The original analysis concerns a monopolist whose decision process is constrained by the imposition of a constraint on the rate of return it can earn on its capital.

Let

q = the firm's output
$R(q)$ = the firm's total revenue function
K = the amount of capital
P_K = the unit price of capital
r = the cost of a dollar's worth of capital
F = the amount of a nonbase input
P_F = the unit price of a nonbase input
s = the regulated rate of return, $s > r$
$G(k, F)$ = the firm's production function.[5]

Then the monopolist's problem is to maximize profits,

$$R(q) - P_F F - rP_K K, \tag{1}$$

subject to the rate of return constraint,

$$R(q) - P_F F - sP_K K \leqslant 0, \tag{2}$$

and subject to the production constraint,

$$q \leqslant G(K, F). \tag{3}$$

The Lagrangian expression for this problem is

$$t = R(q) - P_F F - rP_K K - \lambda[R(q) - P_F F - sP_K K] - \mu[q - G(K, F)]. \tag{4}$$

The first-order conditions are obtained by differentiating expression (4) with respect to q, F, K. Thus we obtain

$$R'(q)(1 - \lambda) - \mu \leqslant 0 \tag{5a}$$

$$-P_F(1 - \lambda) + \mu\, G_F \leqslant 0 \tag{5b}$$

$$-rP_K + \lambda s P_K + \mu\, G_K \leqslant 0 \tag{5c}$$

$$\mu \geqslant 0, \lambda \geqslant 0, q \geqslant 0, K \geqslant 0, F \geqslant 0 \tag{5d}$$

---

[3] See U.S. Federal Power Commission [18].

[4] The interested reader is referred to Baumol and Klevorick [3].

[5] Note that the analysis can account for the presence of more than two inputs: the results are not altered and this formulation is chosen for expositional convenience.

$$\lambda[R(q) - P_F F - sP_K K] = 0 \tag{5e}$$

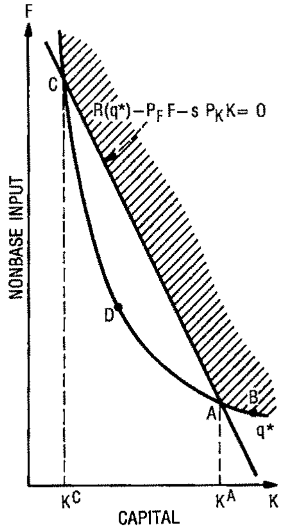
$$\mu[q - G(K, F)] = 0. \tag{5f}$$

All the essential material for the analysis is contained in expressions (5a) through (5f). If we assume this problem has a solution with $q > 0$, then we can replace the inequalities in (5a) through (5c) by equalities. We first observe that $\mu \neq 0$,[6] which implies that the regulated monopoly will operate on its production frontier; thus the regulated firm will not waste resources it has acquired. Also $\lambda \geqslant 0$; from this it can be concluded either that constraint (2) is binding, in which case the firm chooses its inputs in a manner which is inconsistent with economic efficiency, or that constraint (2) is not binding and the firm minimizes its costs of production.[7] But we can go further and indicate the direction of the input inefficiency if we assume strictly convex isoquants. Let q^* be the solution to (5a) through (5f); the rate of return constraint is then a straight line in the (K, F) plane. In Figure 1 the shaded area represents the K-F combinations that satisfy the rate of return constraint. Since q^* is the optimal output level, the set of points to consider is those on the isoquant down to C, then point C, point A, and the set of points on q^* down from A. If the constraint is binding, then the optimal capital is either K^a or K^c; all other admissible capital levels ($K < K^c$, $K > K^a$) are less profitable by the fact that the constraint is binding (otherwise they would have been chosen). It is easy to show that along an isoquant, total costs are increasing as we move away from the input combination yielding minimum cost; therefore, the minimum cost combination is at a point like D; it is easy to discriminate between A and C, since to achieve a given rate of return the point with a higher quantity of capital leads to higher profits. At point D we have

$$\frac{G_K}{G_F} = \frac{rP_K}{P_F}. \tag{6}$$

FIGURE 1

OPTIMAL INPUT MIX



Now since the marginal rate of substitution is decreasing from D towards A and since the ratio of the marginal productivities is equal to the marginal rate of substitution along an isoquant, at point A we have

$$\frac{G_K}{G_F} < \frac{rP_K}{P_K}. \tag{7}$$

Thus, we can conclude that a *monopolist operating under rate of return regulation will either be efficient* (*the regulatory constraint* (2) *is not binding*) *or will overcapitalize* (*constraint* (2) *is binding*).[8] This is the relationship that will be tested later; that is, (6) will be tested against (7).

---

[6] $\mu = 0$ would require $\lambda = 1$ which implies that $s = r$, and the latter was ruled out.

[7] These statements are obtained by substituting for μ in (5b) and (5c) and by dividing the latter by the former.

[8] It should be noted that the proposition just stated has been obtained with very few assumptions. In particular, the traditional Averch-Johnson analysis has always been carried under the assumption that the function

$$R'(K, F) = R(q) \text{ for } q = G(K, F)$$

3. The electric utility industry

The regulation of electric utilities is for most purposes carried out at the state level. At present, 47 State Commissions have jurisdiction over electric companies.[9] In spite of the multiplicity of regulatory bodies, rate-of-return regulation appears to be the essential feature of the regulation of the electric utility industry. This is essentially due to the influence of the U. S. Supreme Court.[10] However, it is recognized that the actual regulatory process is not adequately represented by the traditional Averch-Johnson analysis. Even though electric utilities are granted monopoly rights, regulation is a price-setting mechanism under "fair" rate-of-return criteria.[11] Thus, the regulated firm does not have complete control over its rate structure. However, for a given rate of return there is more than one price structure that can achieve the rate of return. Undoubtedly, a game element as well as dynamic aspects are intrinsic to the regulatory process. The relevant consideration for the purpose of this paper is that the "fair" rate of return is the regulatory tool used to limit (set) prices and that rate of return performance, given the "fair" rate of return, is the method to achieve the end.

The electric utility industry has a complex structure: there exists a wide variety of suppliers ranging from small rural cooperatives usually providing power service for their community exclusively to large corporations distributing power to a huge network of residential and commercial users and to industrial users whose activities depend on electricity. Although they are outnumbered by a ratio of 10 to 1, the privately owned companies account for more than 75 percent of the total supply of delivered electricity.[12] This study will deal exclusively with these companies.

One important characteristic of the industry is the power pool. The demand for electricity varies within days and seasons. Because of the diversity of customers and their geographical location, companies negotiate pooling contracts that enable them to reduce the peak-load pressure on a system or on any individual company. In spite of these pooling arrangements, most companies in the industry are vertically integrated. An electric company is described essentially by three activities if we exclude interests in other fields like gas and transportation which do not involve joint production

---

is concave. There is a strong presumption against this assumption since the traditional rationale for granting monopoly rights subject to regulation has been the alleged presence of economies of scale. Although this motivation for regulation is widely recognized, the analysis of the firm under rate-of-return regulation has been immune to it. The development presented above makes clear that this assumption is not necessary. The assumption of concavity, on the other hand, disallows equilibrium points like B, which at constraint (2) is not binding and is thus characterized by input efficiency.

[9] See Phillips [14].

[10] Three Supreme Court cases are important in this respect. In the *Munn v. Illinois* case (94 U.S. 113-1877), the concept of a public utility was suggested In the *Smyth v. Ames* case (169 U.S. 466-1898) the concept of fair value in the regulation of rates was introduced. It was reinforced in the *FPC v. Hope Natural Gas* case (320 U.S. 591-1944)

[11] For a description of regulatory procedure, the reader is referred to Posner [15].

[12] See Phillips [14].

with electricity in a significant way. These activities are the generation of electricity, its transmission, and its distribution.[13] Electricity is generated by three distinct processes: the combustion of fossil fuels, hydro-power, and nuclear fission. More than 80 percent of the electricity generated in this country comes from the first process and this study will deal exclusively with this type of power plant. Broadly speaking, the generation of electricity in a fossil-fueled plant requires the contribution of three types of input: capital inputs (essentially a furnace-boiler-generator complex), fuel input (coal, gas, and oil, or some combination of these), and labor.[14]

The distinction between transmission and distribution is rather recent and stems from technological advances in the transportation of electricity and changes in population densities. The transmission function is concerned with the transfer of bulk electric loads from a generating station to some distant area. Since the production of electricity requires the presence of some geographical and physical characteristics (mainly the proximity of water sources), transportation of electricity at a high voltage reduces the cost of delivering electricity. The appropriate distinction between distribution and transmission is the voltage. The distribution activity is thus characterized by the delivery of electricity to final users from substations where the voltage of electricity is "stepped down." Both transmission and distribution are characterized almost exclusively by the equipment employed.

In order to evaluate overcapitalization in the electric utility industry, we can consider the firm as an entity or look at each activity separately. However, the first course of action is less desirable, presents serious problems, and is not necessary. It can be argued that the generation and the transmission-distribution activities are separable in the production of delivered electricity. The capital inputs for the transmission and distribution are different from those of the generation with no substitution among them.[15] The same observation would apply for labor input. In other words, the electric utility's production structure can be described by the following production model

$$Q = \text{minimum}\,[G(K, F, L), T(K_t, L_t), D(K_d, L_d)], \qquad (8)$$

where

$G(K, F, L)$ is the production function of the generation activity with K denoting capital, F fuel, and L labor.

$T(K_t, L_t)$ is the production function of the transmission activity with K_t denoting capital input engaged in the transmission of electricity and L_t labor.

---

[13] To obtain an idea of the relative importance of these activities, the *National Power Survey* [18] imputes 51 percent of the total cost of delivered power to generation, 9.9 percent to transmission, and 39.1 percent to distribution.

[14] For a more detailed description, the reader is referred to Cowing [6], Morse [13], or Courville [5].

[15] This is not to say that there are no economic considerations linking the transmission activity and the generation activity. For instance, a power plant needs a cooling agent. The proximity of a water source is usually favored. Higher transmission costs have to be balanced by a more efficient source of cooling and higher fuel and labor costs.

$D(K_d, L_d)$ is the production function of the distribution activity, K_d denoting its capital input and L_d its labor input.

Separability in the production structure of delivered electricity implies that each activity can be considered separately in order to detect possible overcapitalization. Moreover, distribution and transmission are probably characterized by fixed proportions. Then, overcapitalization, if it exists, will occur only in the generation activity. Thus, we can concentrate our attention on this activity exclusively.[16]

There are additional compelling reasons for separating these activities in a study on overcapitalization. A study at the firm level presents serious problems. First, the planning horizon for the transmission and distribution activities is different from the one for the generation of electricity. A distribution system, for instance, is designed and constructed for a greater capacity than the one that is initially used. Without estimates about the range of planning, a study on overcapitalization at the firm level would lead to questionable results. Second, the production of delivered electricity has been marked by technological change. A cross-sectional study at the firm level would have to take account of differences in the technological structure. Finally, some firms' investment in transmission or distribution is higher than others' due to higher participation in pooling power generating facilities, or even more simply,

---

[16] The first proposition in this paragraph is evident. It implies that the firm will bias the use of its base inputs versus nonbase inputs and also that in choosing its base inputs it will follow the same rule as an efficient firm. The second proposition is more interesting, since the implied production model characterizes the production of delivered electricity. It is easy to verify. Let the production structure be defined as:

$$q \leqslant G(K, F, L),$$
$$q = \text{minimum } [\alpha K', \beta L],$$

where

q = output,
K = capital engaged in the first activity,
F = fuel,
L = labor, and
K' = capital engaged in the second activity.

Then the regulated monopolist's problem can be described by the following Lagrangian expression:

$$\begin{aligned} t = R(q) - PF - wL - rP_K K - rP_{K'}K' \\ - \lambda[R(q) - PF - wL - sP_K K - sP_{K'}K'] \\ - \mu_1[q - G(K, F, L)] - \mu_2[q - \alpha K'] - \mu_3[q - \beta L]. \end{aligned}$$

The first-order conditions are:

$$\begin{aligned} & R'(q)(1 - \lambda) - \mu_1 - \mu_2 - \mu_3 = 0 \\ & - P(1 - \lambda) + \mu_1 G_F = 0 \\ & -w(1 - \lambda) + \mu_1 G_L + \mu_3 \beta = 0 \\ & - rP_K + \lambda s P_K + \mu_1 G_K = 0 \\ & - rP_{K'} + \lambda s P_{K'} + \mu_2 \alpha = 0. \end{aligned}$$

From these equations we can infer that $\mu_2 \neq 0$, since $\mu_2 = 0$ implies $\lambda = \frac{r}{s}$. Therefore, no padding will occur and overcapitalization will be inferred only in the activity where substitution is present.

because some firms tend to be net sellers of generated power while others are net buyers.

Therefore, in testing the overcapitalization proposition, we shall consider the power plant as an entity that sells output to the other activities of the utility company. Since overcapitalization is expected to be present in the generation of electricity, we shall consider the objective of the plant manager as minimizing cost with respect to some output level according to appropriate transfer prices. In other words, we view the firm as making the output decision and imputing to the plant manager appropriate relative prices. Since overcapitalization could be optimal from the firm's point of view, the internal price for capital input transferred to the plant manager is less than the external price confronted by the firm as a whole in the capital market.[17]

4. Testing procedure

■ Given an empirical production function for electricity generation, overcapitalization can be tested by comparing the ratio of the marginal productivities to the ratio of prices. The null hypothesis is represented by equation (6) and its alternative by expression (7). The assumption about the relation between the electric utility company and the plant led us to consider input prices as given. Even if the input prices are not known with certainty, minimization of expected cost for a given output level would lead to the same conclusion about the relation between marginal productivities and input prices. Thus, for the purpose of this study, the estimation of a production function relating generated electricity to its factors of production is a crucial step. The theory points to long-run considerations; therefore, an *ex ante* relation, that is, one describing the technical opportunities facing a power plant manager at the planning stage, is appropriate. However, both the acceptance of this concept and its implementation raise some questions.

On the one hand, it can be argued that the *ex ante* production function concept may not be relevant since the firm faces an uncertain future with respect to input prices and output level. Moreover, the firm might consider future capital extensions and there might be a tendency to sacrifice short-run efficiency in order to preserve long-run malleability. Whether or not these phenomena occur and are important remain open questions. In the context of electricity generation, however, these criticisms can be played down somewhat. First, with respect to the uncertainty aspect, the demand for electricity appears to be quite stable along its trend and at the planning stage both capital costs and fuel costs are known, the latter often being determined by long-run contracts. The same thing cannot be said with respect to labor costs. However, with respect to period-by-period changes in input prices, the evidence presented by Fuss[18] suggests that *ex post* substitution has not occurred in electricity generation. Second, with respect to long-run malleability we rely on more tenuous evidence. As Carter has argued,[19] the interaction between new and old capital in the power generating industry appears to be fairly weak. This fact, as Carter also points out, is documented by the Federal Power Commission.[20]

---

[17] It is appropriate at this point to provide additional justifications for the procedure outlined in the previous paragraph, which can be called a "direct approach." Indirect approaches would consist of making use of the first-order conditions directly [(5a)–(5f), above] or through comparative statics so as to relate capital intensity or factor demands to s, the allowed rate of return. There are two sets of problems associated with such procedures. First, knowledge of or assumptions about the revenue function would be necessary. The multiplicity of customers and the peak load phenomenon render this quite difficult in a cross-sectional context. Second, the so-called "fair" rate of return is not easily measurable. On the one hand, there is the problem of "when" to measure. As noted earlier, the Averch-Johnson theory is a long-run description of regulatory behavior which does not incorporate the price-setting (short-run) nature of the regulatory process. On the other hand, there is the problem of "how" to measure. In a cross-sectional context, one would have to be able to account for the observed differences in the treatment of depreciation and rate base valuation methods so as to obtain a more or less homogeneous measure of the allowed rate of return. The direct approach has the advantage of bypassing methodological steps which would depend on additional assumptions.

On the other hand, the implementation of this concept is not free of arbitrariness since, to some extent, it involves the incorporation of expectations. The issue here is how to capture what can be called the "normal" (planned) level of operation. It can be assumed that the year following the construction of the plant usually represents the "normal" year.[21] In the context of multi-plant operation, newer plants which embody recent technology would be given priority in producing output. Thus, even if the actual output of the *firm* falls short of planned output, newer *plants* would receive their allocation first. Thus, for the purposes of estimating the production function, it is appropriate to select what can be justified as representing the normal level of operation. Likewise, it is appropriate to choose the normal level of operation for the purposes of testing and measuring overcapitalization, since it is associated with normal (expected) input consumption.

Without going into details, the data available[22] are yearly observations on the relevant variables. Therefore, the production function to be estimated will relate yearly output to yearly input consumption (fuel and labor) and to capital. However, one of the striking features of electric utility economics is the variation in requested output within the day and within the year. This has an important bearing on the relation to be estimated, since variations within the year in the rate of output affect the performance of the plant.[23] Essentially the consumption of inputs, particularly fuel, is affected by successive changes in the instantaneous rate of output. One way of looking at the problem is to recognize that, for given input consumption, the higher would the annual output be, the lower the variability in output, or that, for a given annual output level, the smaller the input consumption would be, the smaller the variations in output. Thus, variability in output affects factor productivity. Since this study deals with yearly observations and since variations in output occur within the year, cross-sectional differences in the changes in output rate must be accounted for.

---

[18] See Fuss [8].

[19] See Carter [4]. To quote: "Existing steam capacity, with very minor exceptions, has been installed, in fully 'optimized,' that is, balanced units—with boilers and turbo generators paired one-to-one in carefully matched capacity," p. 425.

[20] See U. S. Federal Power Commission [18].

[21] This has been the procedure followed by Dhrymes and Kurz [7] and to some extent by Barzel [2].

[22] An appendix to this paper describing the data is available from the author. The author also wishes to thank Daniel F. McFadden for having supplied the bulk of the data used in this study.

[23] On this point see Morse [13].

5. Production model of electricity generation

■ The production structure in electricity generation has been studied extensively.[24] One of the issues raised by these studies was the specification of a production model.[25] This issue was addressed elsewhere, and the results presented in this paper will be based on the following Cobb-Douglas production model[26]:

$$Q_i = AK_i^{\alpha} F_i^{\beta} L_i^{\lambda} e^{\epsilon_i}, \tag{9}$$

where

Q_i = the output of plant i,
K_i = the capital of plant i,
F_i = fuel consumption of plant i,
L_i = labor consumption of plant i, and
ϵ_i = a random variable normally distributed with mean 0 and variance σ^2 such that $E[\epsilon_i \epsilon_j]_{i \neq j} = 0$.

Several points have to be raised in the context of the empirical production function in electricity generation.

☐ **The measures of capital.** Two different measures of capital are available: a capacity measure and a cost measure. The capacity measure is essentially the maximum output that a given plant can deliver per unit of time: an engineering rating of the plant. The cost measure is the total dollar value of the plant divided into three components: land, structures, and equipment.

The capacity measure has serious defects whose consequences

---

[24] For a review of these studies, the reader is referred to Galatin [12].

[25] See Komiya [10], Barzel [2], Dhrymes and Kurz [7], and Fuss [8]. This issue was spurred by Komiya's failure to discover substitutability among inputs in the generation of electricity.

[26] In Courville [5] two general production models were tested against the Cobb-Douglas model The models postulated belong to the CES family: the first one is the generalized CES, which can be written as:

$$Q = A[\sum_i \alpha_i x_i^{\beta_i}]\gamma, \tag{11a}$$

where Q is output, x_i is input i, and $\Sigma\alpha_i = 1$. This model, estimated with a nonlinear technique, was tested against the usual CES,

$$Q = A[\sum_i \alpha_i x_i^{\beta}]\gamma, \tag{11b}$$

and (11b) was further tested against the Cobb-Douglas specification (11). It should be noted that (11), (11a) and (11b) are tested statistically so that the usual constrained testing procedure applies. In all cases, the Cobb-Douglas formulation could not be rejected. However, it must be recognized that the functional form retained is an approximation and that, to some extent, the results of this study are dependent on the accuracy of the approximation. Also, the specification of the production function as relating expected annual output to expected annual input consumption could suffer from a temporal aggregation problem. To circumvent that problem would require data on an hourly basis.

are difficult to accept. A given level of maximum output per unit of time can be achieved in different ways depending on the relative efficiency of the various capital components. The global efficiency of these components is expressed by what is called the heat rate. Presumably, heat rates will vary across plants for a given unambiguous measure of capacity, because difficult plants face different relative prices. One plant would economize on fuel while the other would economize on the quantity of capital.[27] The cost measure of capital has also its defects which are well-known. It has the disadvantage of confounding price differentials with differences in quality and quantity; its use in production function estimation also confounds quantity discounts with returns to scale.[28] However, it is believed that in the context of electricity generation the capacity measure has more serious defects, since it bypasses the crucial feature of the capital-fuel substitution. Therefore, cost as a measure of capital will be used, but because land does not enter directly in the generation of electricity and also because plant expansion often occurs without any addition to land cost, the cost of land will be excluded.

☐ **Technical change, returns to scale, and size.** The literature on electricity generation indicates that there was significant technical progress over the last thirty years or so—the period partially covered by the sample used in this study; it also indicates the existence of increasing returns to scale. Therefore, if one observes the industry over a period of years, factor substitution, returns to scale, and technical change can be confounded. Throughout this period, the size of plants increased and these plants embodied new technology so that scale and progress have added to capital. Alternatively, technical advances in the capital supplying industry could have resulted in lower prices which led to factor substitution in favor of capital. An attempt to separate out these phenomena will be made by dividing the sample with respect to characteristic vintage years.

The literature on electricity generation also indicates that size, as measured by the capacity rating, is an important variable in the estimation of the production function. Because of limitations on the number of observations, plants of all sizes were grouped together. In order to account for size effects, capacity was introduced as a measure of size in the production relation.[29]

---

[27] Measuring capital by capacity leads to an even more serious problem in the context of an empirical study of overcapitalization. Regulation is a price-setting mechanism and the regulated firm is expected to meet demand at the regulated price. Thus, if capacity measures capital, overcapitalization can be inferred only if excess capacity is present at all points in time.

[28] An additional drawback on the use of cost as a measure of capital is discussed in an appendix available from the author. It relates to the so-called "electric conspiracy" of the late 1950s and the price deflator commonly used for power equipment. Once this is recognized, its impact can be greatly attenuated. The procedure of dividing the sample as described below achieves that end. It should also be mentioned that measuring capital by cost does not remove all measurement errors. The impact of this problem on the results presented in this paper deserves examination.

[29] One observation should be made with respect to size. The treatment

☐ **Output variability.** Since the data on instantaneous output rates are not available, variations in the output rate within the year cannot be measured. However, in the context of electricity generation, the reported figure for average capacity utilization can be used as a proxy for the variation in output rate. This is so because the peak-load output for a given plant is actually equal to the reported capacity (used in computing capacity utilization). *Ceteris paribus*, the smaller the reported capacity utilization, the greater the variation in the output around the mean instantaneous output. Consider the plant as producing two output levels—the peak-load output and the off-peak output; then, the variance in the output rate will be a (decreasing) function of the capacity utilization measure.[30] Finally, it can also be presumed that plants of different capacity do not have the same flexibility in adjusting their output rate; this provides an additional justification for introducing capacity in the relation to be estimated. This relation can now be written as

$$Q_i = AK_i^{\alpha}F_i^{\beta}L_i^{\lambda}e^{\delta u_i}e^{bc_i}e^{\epsilon_i}, \tag{10}$$

where

Q_i, F_i, L_i, ϵ_i are as defined earlier,
u_i = the capacity utilization of plant i, and
c_i = the capacity of plant i.

From production theory and the above discussion, the *a priori* signs of all the coefficients in (10) except b are positive, while the sign of b itself is indeterminate, although it is most likely to be negative.

☐ **Production function results.** The equation fitted for a cross section of plants of a given vintage can be written as:

$$\log Q_i = \log A + \alpha \log K_i + \beta \log F_i + \lambda \log L_i + \delta u_i + bc_i + \epsilon_i. \tag{11}$$

Before proceeding to the presentation of the results, a few comments are in order. The observation chosen on each plant was taken to be the first full year of operation after the plant was constructed. Thus, a plant entered the sample only once; additions to existing plants were not considered. Since on most plants a time series is also available, we tried an alternative procedure to capture the concept of normal operation by taking the average of the logarithms of the yearly output and fuel consumption. The results were less satisfactory. Also, the capacity utilization level enters in a nonlogarithmic form in equation (11). There is no *a priori* reason for that to be: this formulation was retained since it produced better results.[31]

---

of size effects as considered in the text assumes that size affects only A, the scale parameter. Presumably, size can also affect the elasticity parameters α, β, λ. This was investigated and was found not to be the case. See Courville [5]. However, plants for which the initially reported capacity exceeded 800 MW/hour were excluded.

[30] When there are more than two output levels, we can still expect capacity utilization and variability of output rate to be inversely related.

[31] Better results here mean better results from an R^2 point of view. It also happens that when alternative formulations were compared, the one

The results of estimating equation (11) for three different vintage periods are presented in Table 1. The coefficient of labor, $\hat{\lambda}$, had the wrong sign in one case and is not significantly different from zero at the 10-percent level in all three cases. For this reason labor input was excluded from further consideration[32] and the following equation was estimated:

TABLE 1*

REGRESSION ESTIMATES FOR EQUATION (11)

$\text{LOG } Q_i = \text{LOG } A + \alpha \text{ LOG } K_i + \beta \text{ LOG } F_i + \lambda \text{ LOG } L_i + \delta u_i + bc_i + \epsilon_i$

VINTAGE GROUP	$\widehat{\text{LOG } A}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\lambda}$	$\hat{\delta}$	$\hat{b}$	R^2	D F
1948-1950	−1 7779	0 1674	1 0400	−0 006	0 2929	−0 0009	0 995	20
	(2 60)	(2.60)	(16 40)	(0 12)	(1.94)	(2 40)		
1951-1955	−1 5549	0 1254	1 0350	0 0632	0 0869	−0 0003	0 996	30
	(3 54)	(2 99)	(24 71)	(1 62)	(0 88)	(2 62)		
1960-1966	−1 0764	0 0772	0 9493	0 0403	0 2913	0.00009	0 993	31
	(2.66)	(1 59)	(16 53)	(1 03)	(2 47)	(0 62)		

*THE RESULTS PRESENTED IN THIS TABLE ARE BASED UPON USING THE UNDEFLATED MEASURE OF CAPITAL R^2 STANDS FOR R SQUARED ADJUSTED AND D F FOR DEGREES OF FREEDOM IN THE FIRST LINE FOR EACH VINTAGE PERIOD ARE THE VALUES OF THE ESTIMATED COEFFICIENTS THE t-STATISTIC IS THE NUMBER IN PARENTHESES BELOW THE REPORTED VALUES OF THE ESTIMATED COEFFICIENTS

$$\log Q_i = \log A + \alpha \log K_i + \beta \log F_i + \delta u_i + bc_i + \epsilon_i. \quad (12)$$

The results of estimating equation (12) for three different vintage periods are presented in Table 2. Because of the measure of capital chosen and due to the questionable price index available, results based on both the deflated and the undeflated capital measure are presented. The results for the vintage period 1955 to 1959 are presented in Table 3. The estimated coefficients are strikingly different from those presented in Table 2. The coefficient of log K always had the wrong sign and was not significantly different from zero at the 10-percent level in all cases except one. The coefficient of u_i also had the wrong sign. The interpretation of these results in comparison with those of the previous two

---

with the higher R^2 also resulted in a higher estimated value for α, the capital elasticity. This is not a matter indifferent to the purpose of this study.

[32] This point is important and explains why alternative functional forms were estimated; see footnote 26. When a similar test with respect to labor was performed under a CES specification, the same conclusion was reached. This result confirms those of Komiya [10] and of Dhrymes and Kurz [7], while it contradicts the findings of Barzel [2] and Fuss [8]. The difference in the approach followed by the two former studies and this one and by the two latter studies provides an interpretation for this disagreement. In the case of Fuss and Barzel, no separation was made with respect to the vintage years, and their studies used prices of inputs to conclude that labor was substitutable for fuel and capital. Although Dhrymes and Kurz used input prices, they divided their sample into vintage groups as it is done in this paper. The interpretation is that Fuss and Barzel have confounded factor substitution with increasing returns to scale, since the price of labor has increased at the same time that bigger plants were introduced. Dhrymes and Kurz provide evidence on returns to scale to labor.

TABLE 2*

REGRESSION ESTIMATES FOR EQUATION (12)

$\text{LOG } Q_i = \text{LOG } A + \alpha \text{ LOG } K_i + \beta \text{ LOG } F_i + \delta u_i + bc_i + \epsilon_i$

VINTAGE GROUP		$\widehat{\text{LOG A}}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{b}$	R^2	D.F.
1948-1950	U	−1.6255 (2.49)	0.1498 (2.54)	1.0473 (17.10)	0.2625 (1.91)	−0.0009 (2.37)	0.995	23
	D	−0.8715 (2.16)	0.1392 (2.27)	1.0682 (18.11)	0.2281 (1.58)	−0.0010 (2.53)	0.994	23
1951-1955	U	−1.5657 (3.58)	0.1357 (3.31)	1.0042 (22.27)	0.2713 (2.74)	−0.0002 (1.43)	0.994	37
	D	−0.7786 (2.82)	0.1101 (2.43)	1.0269 (21.32)	0.2383 (2.27)	−0.0001 (1.18)	0.994	37
1960-1966	U	−1.2602 (3.40)	0.1036 (3.10)	0.9705 (17.36)	0.3361 (3.04)	0.00002 (0.96)	0.994	35
	D	0.7347 (3.46)	0.1044 (3.08)	0.9711 (17.36)	0.3372 (3.04)	0.00012 (0.13)	0.994	35

*IN THE FIRST COLUMN OF THIS TABLE, THE VINTAGE GROUPS ARE INDICATED A "U" IN THE SECOND COLUMN INDICATES THAT THE MEASURE OF CAPITAL WAS UNDEFLATED WHILE A "D" INDICATES IT WAS DEFLATED.

TABLE 3 *

REGRESSION ESTIMATES FOR EQUATION (12)

$\text{LOG } Q_i = \text{LOG } A + \alpha \text{ LOG } K_i + \beta \text{ LOG } F_i + \delta u_i + bc_i + \epsilon_i$

FOR THE 1956-1959 VINTAGE GROUP

FOR PLANTS WHOSE CAPACITY IS	$\widehat{\text{LOG A}}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{b}$	R^2	D.F.
U <400	0.9603 (2.05)	−0.0667 (1.56)	1.2582 (26.40)	−0.2994 (2.84)	−0.0005 (2.68)	0.998	11
D <400	0.4313 (1.68)	−0.0335 (0.79)	1.2301 (25.86	−0.2475 (2.32)	−0.0005 (2.50)	0.998	11
U <800	0.9873 (2.19)	−0.0719 (1.55)	1.2575 (28.38)	−0.2766 (2.84)	−0.0004 (2.89)	0.999	13
U ALL SIZES	0.8313 (1.70)	−0.0658 (1.52)	1.2034 (26.68)	−0.2383 (2.28)	−0.0001 (3.56)	0.999	18
D ALL SIZES	0.4265 (1.75)	−0.0521 (1.40)	1.1918 (29.27)	−0.2208 (2.19)	−0.0001 (3.53)	0.999	18

*A "U" IN COLUMN 1 INDICATES THAT THE MEASURE OF CAPITAL WAS UNDEFLATED WHILE A "D" INDICATES THAT THIS MEASURE WAS DEFLATED

vintage periods and the following one can be found in an appendix (available from the author). We are forced to conclude that during this period the measurement of capital was subject to error. Basically, the results presented in Table 2 confirm the existence of substitutability between capital and fuel in electricity generation. They also confirm the alleged existence of returns to scale in electricity generation and indicate that there might be returns to scale to fuel alone. In all cases, the coefficient of u_i was significant and had the proper sign.[33]

---

[33] It is understood that the presence of u_i—defined as Q_i/c_i, where c_i stands for capacity—introduces bias in the estimates. However, when u_i

6. Test and measure of overcapitalization

■ In the context of the production function estimated above, it is easy to show that the ratio of the marginal productivities of capital and fuel can be written as

$$\frac{\alpha}{\beta}\frac{F_j}{K_j}, \tag{13}$$

where

K_j = the capital of plant j
F_j = the fuel consumption of plant j.

Then, given the procedure explained earlier,

$$\frac{\alpha}{\beta}\frac{F_j}{K_j} - \frac{P_{K_j}}{P_{F_j}} \tag{14}$$

represents the quantity with respect to which overcapitalization can be inferred. P_{K_j} now represents the (yearly) rental price of capital for plant j, while P_{F_j} is the price of fuel paid by plant j. Therefore, let H_o, the null hypothesis, be

$$\frac{\alpha}{\beta}\frac{F_j}{K_j} - \frac{P_{K_j}}{P_{F_j}} = 0,$$

or equivalently,

$$\frac{\alpha F_j}{K_j} - \frac{\beta P_{K_j}}{P_{F_j}} = 0. \tag{15}$$

Moreover, let

$$T^*_j = \frac{\dfrac{\hat{\alpha} F_j}{K_j} - \dfrac{\hat{\beta} P_{K_j}}{P_{F_j}}}{\sqrt{\sigma^2}}, \tag{16}$$

where

$$\sigma^2 = \sigma_{\hat{\alpha}}^2 \left(\frac{F_j}{K_j}\right)^2 + \sigma_{\hat{\beta}}^2 \left(\frac{P_{K_j}}{P_{F_j}}\right)^2 - 2\sigma_{\hat{\alpha}\hat{\beta}} \frac{F_j}{K_j} \cdot \frac{P_{K_j}}{P_{F_j}}$$

$\sigma_{\hat{\alpha}}^2$ = estimated variance of $\hat{\alpha}$
$\sigma_{\hat{\beta}}^2$ = estimated variance of $\hat{\beta}$
$\sigma_{\hat{\alpha}\hat{\beta}}$ = estimated covariance between $\hat{\alpha}$ and $\hat{\beta}$, where $\hat{\alpha}$ and $\hat{\beta}$ are the linear regression estimates of α and β.

If F_j and K_j are considered nonstochastic, then, under the null hypothesis, T_j^* has a Student t-distribution with $n - k$ degrees of freedom, where n is the number of observations and k is the number of parameters estimated in the regression. The region of rejection consists of negative and low values of T^*. Rejection of H_o confirms overcapitalization, since rejection of H_o with low values of T^* also leads to the rejection of

$$\frac{\alpha F_j}{\beta K_j} - \frac{P_{K_j}}{P_{F_j}} > 0.$$

---

was taken to be the average of the yearly utilization rates excluding the first year for a given plant, thus implying no bias, the estimated value of α was smaller. For the purpose of testing for overcapitalization, the procedure we have chosen works against finding overcapitalization.

In performing the test for each plant in different samples, P_{K_j} was obtained by using the average yield to maturity of AAA public utility bonds the year preceding the first full year of operation, as computed by Moody's. In order to account for depreciation, an estimate of the average life of electric utility plant was used. The estimate taken was forty years, and this corresponds to the one used by Livingstone.[34] Then, the capital recovery factor was computed, that is, the uniform annual payment for an annuity of forty years, which is equivalent (at a rate of interest equal to the AAA yield) to one dollar of initial capital outlay. This is the annual cost of one dollar of capital held for forty years assuming zero salvage value.[35] The annual cost per dollar (the capital recovery factor) was then multiplied by the total dollar value of the plant to get P_{K_j}. When the deflated value for the measure of capital was used, P_{K_j} was further multiplied by the value of the deflator for that year.

The price of fuel, P_{F_j}, was taken to be the average over the years for which observations were available on each plant. This procedure for computing expected fuel prices assumes static expectations about the distribution of future levels of prices. This is probably appropriate, since the price a firm pays for fuel over those years is quite stable.[36] The main reason for this lies in the fact that fuel contracts between the suppliers and the electric utility company are typically long-term contracts.

Tables 4 and 5 provide a summary of the results on overcapitalization. Due to space limitation—there were 110 plants in the three vintage periods retained—the t-values for each plant are not presented. Instead, the summary indicates the number of times H_0 was rejected and some information about the value of the T_j^*'s. The only difference between Table 4 and Table 5 is that the overcapitalization tests presented in Table 4 correspond to the empirical production function estimated using the undeflated measure of capital, while in Table 5 the tests are based on using the deflated measure of capital.

The first column of each table summarizes the values of T_j^* when F_j in equation (15) was taken to be the fuel consumption for the first year of operation for plant j. Given the formulation of the *ex ante* process and the related decision model facing the

---

[34] See Livingstone [11].

[35] The yield to maturity of public utility bonds is not the appropriate rate to be used in calculating these capital costs, but it can be considered as a lower bound. In their computations made of the cost of capital to the electric utility industry for the years 1954, 1956, and 1957, Miller and Modigliani [12] found that the yield on the AAA bonds was lower than their estimates of the cost of capital to electric utility firms. Since the ratio of debt to total assets is fairly stable in this industry, it can be stated that this relationship between AAA bond yield and the cost of capital is valid for the years covered by the sample used in this study, so that as far as the electric utility industry is concerned, the AAA bond yield underestimates the cost of capital to the industry. Also, this procedure does not allow for capital cost differences across firms; thus, choosing the better rated bonds establishes a lower bound for cost of capital for each firm in the sample.

[36] Static expectations with respect to fuel prices have also been used by Fuss [8].

TABLE 4

SUMMARY TABLE FOR THE OVERCAPITALIZATION TESTS (16)
USING THE UNDEFLATED MEASURE OF CAPITAL

TOTAL NUMBER OF PLANTS	110	110	110
NUMBER OF TIMES H_0 WAS REJECTED *AT 5% LEVEL	105	99	71
AT 10% LEVEL	107	104	76
RANGE OF VALUES OF T^*_j	−0 33, −12 13	−0 21, −11 2	+0 35, −8 15

*USING A ONE-TAIL TEST

TABLE 5

SUMMARY TABLE FOR THE OVERCAPITALIZATION TESTS (16)
USING THE DEFLATED MEASURE OF CAPITAL

TOTAL NUMBER OF PLANTS	110	110	110
NUMBER OF TIMES H_0 WAS REJECTED. *AT 5% LEVEL	107	104	77
AT 10% LEVEL	108	106	85
RANGE OF VALUES OF T^*_j	−0 52, −11 72	−0 41, −11.27	−10 09, −8 17

*USING A ONE-TAIL TEST

plant manager, this test is the appropriate one. However, it is recognized that the first full year of operation may not represent fully anticipated production level and fuel expenses. This is of some importance in the context of the electric utility industry, since plants operate at different output levels throughout the year. It is also recognized that the production function retained is not the exact form, but an approximation. For these reasons procedures biasing the test against the rejection of the null hypothesis were attempted. The first procedure consists of taking as F_j, in equation (15), the fuel consumption for plant j at the highest yearly output level encountered for this plant. The results of the overcapitalization tests using fuel consumption at the highest annual output are summarized in column 2 of Tables 4 and 5. The second procedure imposes an even greater bias. The highest yearly output does not represent the highest output generated by the plant, since it does not correspond to the peak-output.[37] In order to account for this, an estimate of fuel consumption $\hat{F}_j$ at peak-output—as measured by plant capacity—was obtained by using the estimated production function.[38] However, in computing the T_j^*'s, as summarized in column 3 of Tables 4 and 5, $\hat{F}_j$ was considered nonstochastic.[39]

---

[37] A cost-minimizer, faced with varying demand levels and producing a nonstorable commodity, will choose its inputs in such a way as to minimize the total costs of operation. For the sake of exposition, assume two output levels: off-peak and peak-load. Thus, at peak output, a cost-minimizer would undercapitalize to a degree depending on the relative frequency of operation at peak output. That is, if the consumption of inputs of a cost-minimizer is observed at peak output, then undercapitalization should be inferred. It should also be recognized that the production function estimated relates planned output and planned input consumption. For the biasing procedures to be valid, the assumption that this relation is an adequate approximation of the relation between output and inputs at other levels than planned output is required.

[38] Given an estimate of the production function, one can obtain an estimate of the fuel that would be used at peak output. Let $\hat{F}_j$ be that esti-

The results presented in Tables 4 and 5 confirm the existence of overcapitalization in the electric utility industry. The Averch-Johnson analysis is an abstract but simple description of the behavior of firms under rate-of-return regulation and the data on the electric utility industry support the hypothesis. The results presented are consistent with those of Spann,[40] who found overcapitalization in the electric utility industry, and they also confirm a number of casual observations.[41]

There is still a question as to the importance of the departures of K/F from the minimum cost combination. While the results discussed so far confirm the existence of overcapitalization, they do not indicate the economic significance of this phenomenon. One would like to have an estimate of the annual dollar costs implied by inefficient production of electricity. To shed some light on this issue, the percentage deviations of actual cost from minimum cost have been computed and information about the magnitude of the deviations is presented in Table 6. In order to obtain the value of the percentage deviation from minimum cost, the optimal (from a cost-minimizing point of view) values for fuel and capital were computed by using the first-order condition together with the em-

TABLE 6

SUMMARY TABLE FOR THE PERCENTAGE DEVIATIONS OF MINIMUM COST FROM ACTUAL COST

	AT ACTUAL OUTPUT	AT PREDICTED OUTPUT	AT PEAK OUTPUT
AVERAGE	11.42	11.95	6.01
RANGE	-0.6, 40.6	0.1, 42.7	0.1, 25.8
NUMBER OF TIMES LESS THAN AVERAGE	62	61	41
TOTAL NUMBER OF PLANTS	105	105	71

---

mate; it is computed by solving for $\widehat{\log F_j}$ in the empirical production function and then by letting

$$\hat{F}_j = e^{\widehat{\log F_j}}.$$

$\hat{F}_j$ is a consistent estimator, but it is biased, since

$$E[F_j] = E(e^{\widehat{\log F_j}}) > e^{E(\widehat{\log F_j})} = e^{\log F_j} = Fj.$$

Thus, this last procedure is biased in favor of the null hypothesis on two counts.

[39] $\hat{X} = \dfrac{Y_0 - \hat{a}}{\hat{b}}$ in the regression model $Y_i = a + bx_i + u_i$ is known to have an infinite mean-squared error. Thus, distribution properties of $\hat{F}_j$ are intractable.

[40] In [16].

[41] See in particular Averch and Johnson [1].

pirical production function. That is, the following system of equations was solved for F_j and K_j:

$$F_j = \frac{\hat{\beta}}{\hat{\alpha}} K_j \frac{P_{K_j}}{P_{F_j}} \qquad (17)$$

$$\log Q_j = \widehat{\log A} + \hat{\alpha} \log K_j + \hat{\beta} \log F_j + \hat{\delta} u_j + \hat{b} c_j.^{42} \qquad (18)$$

Denote by F_j^* and K_j^*, the solution to these equations. Then, Δ_j, the percentage deviation from the minimum cost combinations, can be expressed as

$$\Delta_j = \frac{P_{F_j} F_j^0 + P_{K_j} K_j^0 - P_{F_j} F_j^* + P_{K_j} K_j^*}{P_{F_j} F_j^0 + P_{K_j} K_j^0},$$

where $P_{F_j} F_j^0 + P_{K_j} K_j^0$ represents actual cost and $P_{F_j} F_j^* + P_{K_j} K_j^*$ represents minimum cost.[43] The average of the percentage deviations from minimum as well as other relevant information is presented in column 1 of Table 6. It should be mentioned that differences between actual and predicted output sometimes resulted in a negative value for Δ_j; this occurred in six out of a possible 105 cases. The information summarized in column 2 of Table 6 was obtained by performing the same exercise, except that this time predicted output was used [in equation (18)] to infer the value of the minimum cost combination.[44] This procedure is more desirable, since it represents the problem a cost-minimizer would face in the presence of uncertainty in the production process. Finally, a last set of numbers was generated so as to obtain a lower bound on the economic importance of inefficient electricity generation. For the percentage deviations from minimum summarized in the third column of Table 6, $\hat{F}_j$[45] was used. This represents the deviation from minimum cost that would obtain if plants were operating at peak output all the time. Actual cost in this case was

---

[42] The regression based on the undeflated measure of capital was used for equation (18).

[43] A completely satisfactory answer to the computation of the percentage deviation from minimum cost is difficult to obtain. Again, it should be noted that the empirical production function relates yearly values of the relevant variables and abstracts from operation at various output levels. Statistical difficulties are also encountered. First, the inverse estimation problem mentioned earlier—see note 39—is still present. Second, the implied value of the minimum cost for plant j will be a function of the estimated residual j in the production function regression. However, note that insofar as the regression estimates are consistent, $P_{F_j} F^*_j + P_{K_j} K^*_j$ is also a consistent estimate of the minimum cost. Another problem is also encountered and it has to do with the particular features of regulation. Because regulation is a price-setting mechanism, regulated firms are requested to meet demand at the regulated price. Thus, to some extent, the output that the firm will have to deliver is given and in particular so is the peak-load output. This constraint gets reflected at the plant level. The computation of the efficient input levels (from the first-order conditions) implicitly assumes that the resultant figure for capital is a feasible solution with respect to ("given") peak output; that is, it is assumed that the optimal amount of dollars can buy equipment providing the same capacity output that the plant actually provides.

[44] It is easy to see that in this case the value of minimum cost is not dependent on $\widehat{\log A}$, $\hat{\delta}$, and $\hat{b}$.

[45] See note 38.

measured by $P_{K_j} K_j + P_{F_j} \hat{F}_j$ and minimum cost was inferred as previously by using reported capacity as the measure of peak output.

When comparing the results presented in Table 6 with the values of the overcapitalization tests presented in Tables 4 and 5, one cannot escape noticing that in spite of the fact that overcapitalization was strongly confirmed, the actual deviation from minimum cost is relatively small.

The National Power Survey[46] estimated that in 1962 the average total production cost of conventional steam-electric power produced by privately owned utilities was slightly above 7.5 mills per kilowatt-hour. Operation and maintenance (essentially labor) account for 0.79 mills and fuel for 2.75 mills. This leaves an imputed 3.96 mills for total capital. Land in 1962 accounted for 1.1 percent of total plant, which would leave an imputed cost of 3.91 mills for capital costs, exclusive of land, per kilowatt-hour. Thus, the total cost (fuel + capital excluding land) of providing electricity is 6.6 mills per kilowatt-hour. In 1962, privately owned electric utilities sold a total of 685.7 billion kilowatt-hours[47] of which approximately 80 percent[48] was produced by conventional steam-electric power plants or 548.6 billion kilowatt-hours. Thus the total cost of generating this quantity of electricity can be estimated at 6.66 times 548.6, or 3.653 billion dollars. By using the estimate (11.95) of the deviation from minimum cost, the total saving that would have occurred in 1962 from an efficient production of electricity by steam power plants is estimated at 436.5 million dollars[49] (if all plants producing electricity in 1962 were inefficient to a degree comparable to the ones in the sample).

7. Concluding comments

■ The Averch-Johnson proposition attempts to describe the behavior of monopolists under rate of return regulation. The aim of this study has been to quantify this proposition with respect to the electric utility industry. The results are consistent with the hypothesis that this form of regulation induces overcapitalization. These results have been obtained by estimating a production function for electricity generation and by comparing the implied ratio of the marginal productivities of capital and fuel to the ratio of their respective prices. It was argued that this procedure is the only effective way to quantify overcapitalization. It also allowed for devising biases against finding overcapitalization.

---

[46] See FPC [18].

[47] See FPC [19].

[48] See FPC [18].

[49] With the lower bound estimate for the percentage deviation from minimum cost, the total savings due to efficient production would have been 219.5 million dollars. Given the methodology adopted in this paper, this lower bound does not have much significance. However, the procedure for obtaining biased estimates for both the test and measure of overcapitalization is motivated by the recognition that the concept of normal operation presents measurement problems and captures only implicitly operation at various output levels. Thus, in that sense, it can be said that the economic gain to society from potential efficient production can be estimated at between 219.5 and 436.5 million dollars.

A measure of the importance of the inefficiency in producing electricity was obtained. Even though the percentage deviation of actual cost from minimum cost was not very large for most of the individual plants in the sample used in this study, the implied dollar value for the total electricity generated by steam electric power plants is not negligible. For the year 1962 alone, it was estimated that efficient production of electricity would have resulted in a saving estimated at 436.5 million dollars or 11.95 percent of the total actual production costs.

It should be recognized and emphasized that this study *does not* represent an evaluation of the regulatory process in the electric utility industry. Presumably, regulation induces benefits by forcing firms to operate at a greater output than the pure monopoly output. Thus, in principle, one should compute the gains to consumers due to regulation and compare them to the increased costs generated by inefficient production induced by rate-of-return regulation. It is well-known that such a comparison may not be completely satisfactory, since this exercise looks at a particular institutional environment and presumes that the economic relations will stay the same under alternative arrangements. However, the results of this study indicate the desirability of such an investigation, especially when they are contrasted with the conclusions of Stigler and Friedland,[50] who have argued that the price of electricity had not been significantly affected by the introduction of regulation in its early days. This last observation provides a strong motivation for carrying out a comparison of the costs and benefits of regulation under its present arrangements, since Stigler and Friedland's findings, in conjunction with the results of the present paper, could mean that regulation has induced the worst of all worlds.

References

1. Averch, H. and Johnson, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. Barzel, Y. "The Production Function and Technical Change in the Steam Power Industry." *Journal of Political Economy*, Vol. 72, No. 2 (April 1964), pp. 130–150.
3. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
4. Carter, A. P. "Comment" in *The Theory and Empirical Analysis of Production*, National Bureau of Economic Research, Studies in Income and Wealth, No. 31, Brown, M., ed., New York, 1967.
5. Courville, L. "Regulation and Efficiency in the Electric Utility Industry." Dissertation to be presented to the Faculty of the Graduate School of Industrial Administration, Carnegie-Mellon University.
6. Cowing, T. G. "Engineering Process Analysis and Neoclassical Production Functions: A Reconciliation." Paper delivered at the December 1970 meetings of the Econometric Society.
7. Dhrymes, P. J. and Kurz, M. "Technology and Scale in Electricity Generation." *Econometrica*, Vol. 32, No. 4 (July 1964), pp. 287–315.

---

[50] See Stigler and Friedland [17].

8. Fuss, M. A. "Factor Substitution in Electricity Generation: A Test of the Putty-Clay Hypothesis." Harvard Institute of Economic Research, Discussion Paper Number 185, 1971.
9. Galatin, M. "Economies of Scale and Technological Change in Thermal Power Generation." Amsterdam: North-Holland Publishing Co., 1968.
10. Komiya, R. "Technological Progress and the Production Function in the United States Steam Power Industry." *Review of Economics and Statistics*, Vol. 44, No. 2 (May 1962), pp. 156–166.
11. Livingstone, J. L. "Accelerated Depreciation, Cyclical Asset Expenditures and Deferred Taxes." *Journal of Accounting Research*, Vol. 5, No. 1 (Spring 1967), pp. 77–94.
12. Miller, M. H. and Modigliani, F. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–1957." *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 333–391.
13. Morse, F. T. *Power Plant Engineering*. New York: D. Van Nostrand Co., 1953.
14. Phillips, C. F. *The Economics of Regulation*. Homewood, Ill.: R. D. Irwin, 1965.
15. Posner, R. A. "Natural Monopoly and Its Regulation." *Stanford Law Review*, Vol. 21, No. 3 (1969), pp. 548–643.
16. Spann, R. M. "Rate of Return Regulation and Efficiency in Production: An Empirical Test of the Averch-Johnson Thesis." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 1 (Spring 1974), pp. 38–52.
17. Stigler, G. J. and Friedland, C. "What Can Regulators Regulate? The Case of Electricity." *Journal of Law and Economics*, Vol. 5 (1962), pp. 1–16.
18. U. S. Federal Power Commission. *National Power Survey*, Vol. 1. Washington, D. C.: U. S. Government Printing Office, 1964.
19. ———. *Statistics of Electric Utilities, Classes A and B Privately Owned Companies*. Washington, D. C.: U. S. Government Printing Office, 1969.
20. ———. *Steam Electric Plants: Construction Costs and Annual Production Expenses, 1948–69*. Washington, D. C.: U. S. Government Printing Office, annually.

Reversals in peak and offpeak prices

Elizabeth E. Bailey

Supervisor—Economic Analysis
Bell Telephone Laboratories

Adjunct Assistant Professor of Economics
New York University

and

Lawrence J. White

Assistant Professor of Economics
Princeton University

This paper examines the pattern of peak and offpeak prices for several models of firm behavior beyond the standard welfare-maximizing models of Boiteux, Steiner, and Williamson. In the case where there is a profit objective or a breakeven constraint, we show that it can be rational for the firm to set a higher price in the offpeak than in the peak period; indeed, this occurs whenever offpeak demand is sufficiently more inelastic than peak demand so as to compensate at the margin for the attribution of capacity costs to the peak user. Under regulation the price reversal takes place because the regulated firm best serves its owners if it lowers prices only to peak users. Further, the paper demonstrates the possibility that rate-of-return regulation could induce peak capacity which is greater than the socially optimal level. This last point may have relevance for present and future debates over the amounts of new capacity in energy-generating industries that ought to be constructed to meet the "needs" of the economy.

1. Introduction

■ Can it be rational for the supplier of a monopoly service, such as transportation or public utilities, to charge a price during peak periods that is *lower* than the price in offpeak periods? Such a pricing

---

Elizabeth E. Bailey received the A. B. in economics from Radcliffe College (1960), the M. S. in mathematics and computer science from Stevens Institute of Technology (1966), and the Ph. D. in economics from Princeton University (1972). She is Supervisor—Economic Analysis, Bell Telephone Laboratories, Holmdel, New Jersey, and Adjunct Assistant Professor of Economics at New York University. Her current research interests are in pricing policies for regulated industries, and she is also working on a model which shows the impact of volunteer activities on the economy.

Lawrence J. White received the A. B. from Harvard (1964), the M. Sc. (Econ.) from the London School of Economics (1965), and the Ph. D. from Harvard (1969). As Assistant Professor of Economics at Princeton, his current research is in microeconomic models that include quality variables, models that relate domestic industrial structure to international trade flows, and a study of industrial concentration and economic power in Pakistan.

possibility has been virtually ignored in economists' discussions of peak-load pricing. Yet, such pricing structures sometimes are found in practice. For instance, the daily peak in an electric utility is often generated mainly by business customers who pay less for service than residential customers whose evening demand is offpeak.

Most theoretical treatments of peak-load pricing reach the conclusion that peak prices will be higher than offpeak prices. This result follows automatically in the standard model in which there is a welfare objective and the firm has constant returns to scale in production. We show that a higher peak price also occurs in a model in which the firm can levy a perfect two-part tariff.

What has not been recognized is that a peak price less than the offpeak price *can be optimal* in other models of firm behavior. Specifically, reversals can occur when:

(1) The firm has a profit rather than a welfare objective, and faces a demand in the offpeak period that is sufficiently more *inelastic* than that during the peak so as to compensate at the margin for the attribution of capacity costs to the peak period;
(2) The firm is subject to a rate-of-return constraint, which gives it an incentive to lower the rate to the peak users while keeping the offpeak rate at the monopoly level;
(3) The firm has a welfare objective, but there are decreasing average costs in production, so that a profit constraint is explicitly or implicitly imposed; and
(4) The firm employs a multipart pricing scheme, but one that is imperfect in extracting all the consumer's surplus.

Such models do not exhaust the situations that can lead to pricing reversals, but they do suggest that there are a number of important cases in which such reversals can be explained by rational economic processes.

The models are also important in a second respect, for in some of them, we are able to demonstrate that capacity may be expanded beyond the level that maximizes welfare. In particular, rate-of-return regulation can lead to such overexpansion of capacity, and may do so whether a single or a multipart tariff is levied.

2. The profit versus the welfare objective

■ **Welfare maximization.** Virtually all of the literature dealing with peak-load pricing principles seeks to determine the conditions necessary for maximization of social welfare in the sense of Pareto

---

We are particularly grateful to Paul W. MacAvoy, William J. Baumol, Gerald R. Faulhaber and Edward E. Zajac, each of whom made substantial contributions to the paper. Our thanks also go to a number of other colleagues in the Bell System and in the Academic Community who were of help to us. Part of this research was conducted while Lawrence White held a Post-Doctoral Fellowship in Public Utility Economics from the American Telephone and Telegraph Company. The views expressed in the paper are those of the authors, and should not be interpreted as reflecting official opinion or policy of Bell Laboratories or of the American Telephone and Telegraph Company.

optimality or of maximization of consumer's plus producer's surplus.[1] We shall begin by reviewing the results obtained with a simple version of this model.

Our presentation assumes that there are two periods of equal duration and that demands in the two periods are not related. To make this seem sensible, think of the model as depicting two classes of customers, such as business users of power and residential users, whose demand for the facilities occurs during the day and evening, respectively.[2] The cost structure consists of a constant operating expense per unit of demand served, and a constant cost per unit of capacity. We shall suppose that the outcome in terms of capacity usage is that there is a distinct period of peak and a period of offpeak (rather than that there is a joint peak). These simplifying assumptions are made so that we can focus directly on the pricing reversal phenomenon, and so that straightforward graphical interpretations can be presented. The consequences of relaxing some of the assumptions are discussed later in the paper.

The mathematical formulation portrays the maximization of the sum of producer's and consumer's surplus, given by the integrals of demand curves less costs, as summarized by model (1) in Table 1, where $\bar{X}$ and $\underline{X}$ are demands in the peak ($\bar{X}$) and offpeak ($\underline{X}$) periods, respectively, with peak period demand equaling capacity, $\bar{p}$ and $\underline{p}$ are prices in the peak ($\bar{p}$) and offpeak ($\underline{p}$) periods, respectively, c is the operating expense per unit of output (constant), k is the acquisition cost per unit of capacity (constant), r is the market cost of capital (constant), and the subscript W denotes values relating to the welfare-maximization objective.

Equations (2) and (3) indicate that peak price covers both the expense cost generated by the peak traffic and the marginal capacity cost, whereas offpeak price just covers its own expense cost. It is immediately obvious from (2) and (3) that

$$\bar{p}_W > \underline{p}_W,$$

so that *if there are constant costs, welfare maximization automatically requires the peak price to be higher than the offpeak price.* The familiar pricing rule has appeared.

The results are made clearer by Steiner's peak-load geometry[3] as displayed in Figure 1. Equation (2) is just satisfied at point $\overline{W}$. Equation (3) is satisfied at $\underline{W}$. Since price is equated with the

---

[1] See, especially, Boiteux [7], Steiner [26, 27], Hirshleifer [15], Wellisz [30], Williamson [31], Turvey [28], Littlechild [18], Mohring [20], and Pressman [24].

[2] In essence, we have a model of price discrimination in which demands are sequential rather than simultaneous. For a suggestive daily load curve showing that the electricity example fits this description in at least an approximative manner, see *Electric Power and Government Policy* [13], p. 38. In practice, peaks in the electric utilities are attributable, at least in part, to both business and to residential customers; however, the different classes of customers are likely to have different peak intensities.

[3] In [27].

TABLE 1

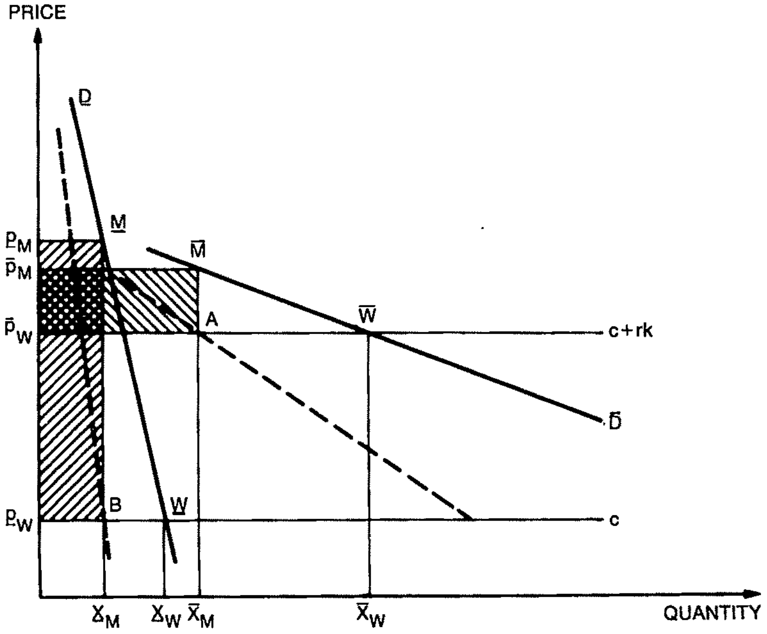
SUMMARY OF RESULTS

MODEL	PEAK PRICE	OFFPEAK PRICE
WELFARE MAXIMIZATION (1) $\underset{\bar{X},\underline{X}}{\text{MAXIMIZE}}\ W = \int_0^{\bar{X}} \bar{p}\, d\bar{X} + \int_0^{\underline{X}} \underline{p}\, d\underline{X} - c\,(\bar{X}+\underline{X}) - rk\bar{X}$	(2) $\bar{p}_W = c + rk$	(3) $\underline{p}_W = c$
PROFIT MAXIMIZATION (4) $\underset{\bar{X},\underline{X}}{\text{MAXIMIZE}}\ \pi = \bar{p}(\bar{X})\bar{X} + \underline{p}(\underline{X})\underline{X} - c(\bar{X}+\underline{X}) - rk\bar{X}$	(5) $\bar{p}_M = \dfrac{c+rk}{1-\frac{1}{\bar{e}}}$	(6) $\underline{p}_M = \dfrac{c}{1-\frac{1}{\underline{e}}}$
INCREASING RETURNS IN CAPACITY PROVISION (7) $\underset{\bar{X},\underline{X}}{\text{MAXIMIZE}}\ W = \int_0^{\bar{X}} \bar{p}\, d\bar{X} + \int_0^{\underline{X}} \underline{p}\, d\underline{X} - c(\bar{X}+\underline{X}) - K(\bar{X})$ SUBJECT TO $\bar{p}(\bar{X})\bar{X} + \underline{p}(\underline{X})\underline{X} - c\,(\bar{X}+\underline{X}) - K(\bar{X}) \geqslant 0$	(8) $\bar{p}_I = \dfrac{(1+\lambda)(c+K')}{1+\lambda\,(1-\frac{1}{\bar{e}})}$	(9) $\underline{p}_I = \dfrac{(1+\lambda)\,c}{1+\lambda\,(1-\frac{1}{\underline{e}})}$
RATE-OF-RETURN REGULATION (10) $\underset{\bar{X},\underline{X},Z}{\text{MAXIMIZE}}\ \pi = \bar{p}(\bar{X})\bar{X} + \underline{p}(\underline{X})\underline{X} - c(\bar{X}+\underline{X}) - rk(\bar{X}+Z)$ SUBJECT TO $\bar{p}(\bar{X})\bar{X} + \underline{p}(\underline{X})\underline{X} - c(\bar{X}+\underline{X}) - sk(\bar{X}+Z) \leqslant 0$	(11) $\bar{p}_G = \dfrac{c + rk - \frac{\lambda}{1-\lambda}(s-r)k}{1-\frac{1}{\bar{e}}}$	(12) $\underline{p}_G = \dfrac{c}{1-\frac{1}{\underline{e}}}$
PERFECT TWO-PART TARIFF, SEE TEXT	(23) $\bar{p}_T = c + rk$	(24) $\underline{p}_T = c$
REGULATION OF RATE OF RETURN BUT NOT OF PRICE STRUCTURE, SEE TEXT	(25) $\bar{p}_R = c + rk - \frac{\lambda}{1-\lambda}\,(s-r)k$	(26) $\underline{p}_R = c$

relevant costs in each case, and since costs are constant, the firm just exactly breaks even by supplying capacity $X = \bar{X}_W$ and total demand $D = \bar{X}_W + \underline{X}_W$ at prices $\bar{p}_W$ and $\underline{p}_W$, respectively.

FIGURE 1

WELFARE VERSUS MONOPOLY PRICING



☐ **Profit maximization.** Now let us contrast these results with those that follow when the firm's objective is to maximize profit rather than producer's plus consumer's surplus. The relevant model is (4) in Table 1, where the subscript M is used to denote values relating to the profit objective of the monopoly firm. The $\bar{e}$ and $\underline{e}$ are price elasticities of demand in the peak and offpeak periods, respectively.

When elasticities are equal, (5) and (6) show that peak and offpeak prices are in the same ratio as the marginal costs of peak and offpeak service, and thus offpeak price in this case will always be lower than peak price. However, offpeak rates are *higher* than peak rates whenever

$$c\left[\frac{1}{1-\frac{1}{\underline{e}}}\right] > (c+rk)\left[\frac{1}{1-\frac{1}{\bar{e}}}\right]$$

Thus, *offpeak price can be higher than peak price if offpeak service has demand that is sufficiently more inelastic than the peak service so as to compensate for the larger marginal costs attributable to the peak service. It is clear that the size of rk relative to c is also important in the determination of pricing reversals. If marginal capacity costs are large relative to marginal operating expenses, then the pricing reversal is less likely to occur.*

In terms of the electricity example, the pricing reversal may occur if rate schedules are "a result of pricing policies based on value of service."[4] Residential users place a higher value on such service (i.e., have a more inelastic demand) than do the business users, since large business customers are more likely to be able to generate their own electricity or to convert to alternative sources of energy.

The monopoly solution is described graphically in Figure 1. Equations (5) and (6) are satisfied at points $\bar{M}$ and $\underline{M}$ respectively, with $\underline{p}_M > \bar{p}_M$ so that the offpeak price is higher and the reversal occurs. Profits are given by the sum of the two striped rectangles $\underline{p}_M \underline{M} B \underline{p}_W$ and $\bar{p}_M \bar{M} A \bar{p}_W$, with the markup of offpeak price over operating expense being relatively larger than the markup of peak price over the sum of operating and capacity expense. Notice that it is not correct to refer to rectangle $\underline{p}_M \underline{M} B \underline{p}_W$ as "offpeak profit" and $\bar{p}_M \bar{M} A \bar{p}_W$ as "peak profit." Since peak and offpeak users share some common capacity, any attribution of costs and profits runs into the familiar joint-cost problem.

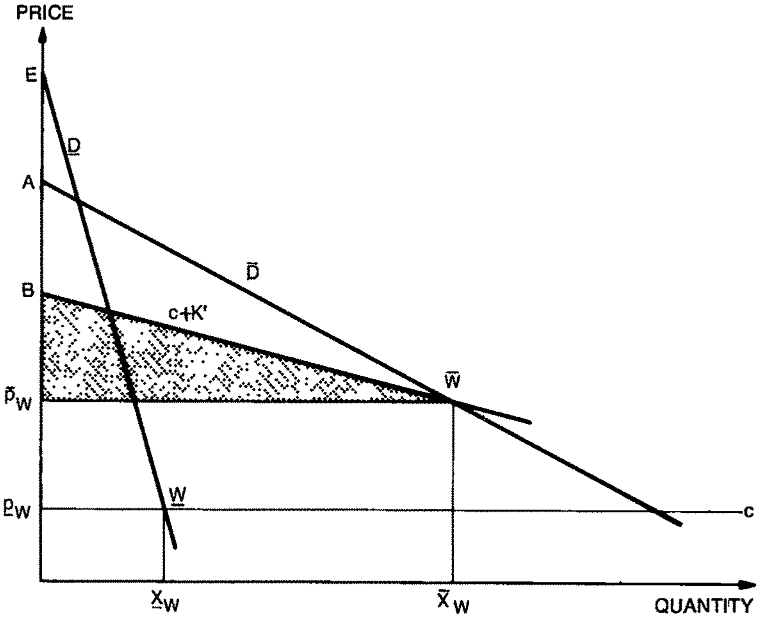
☐ **Welfare maximization with increasing returns in capacity provision.** Many firms for which the peak-load pricing problem is relevant have increasing returns to scale in the provision of capacity,[5] that is, capacity costs are given by $K(\overline{X})$, $K' > 0$, $K'' < 0$. Figure 2 illustrates the situation. At the output pair, $(\overline{W}, \underline{W})$ where welfare is maximized, price equals marginal cost for each user, $\underline{p}_W = c$ and $\overline{p}_W = c + K'$. Under such a scheme the firm operates at a loss whose magnitude is given by the dotted area $B\overline{W}\overline{p}_W$. If a single price is all that can be levied for each type of customer, and if the firm is to break even in its operations, the appropriate model is given by (7) in Table 1.

---

[4] Davidson [12], p. 215. Notice also that the pricing reversal might be explained if the cost of distributing electricity to the residential customers (the c term) were higher than the distribution cost to the business customer. See, for example, MacAvoy and Noll [19]. Another complication that can arise in the real situation is that the peak and offpeak periods of the generating and distribution systems might occur at different times.

FIGURE 2

INCREASING RETURNS TO CAPACITY



The pricing rules (8) and (9) can be interpreted as stating that the deviation of price from marginal cost for each period should be proportional to the marginal deficit incurred by the last unit of output in that period.[6] It is clear that we can obtain from (8) and (9) an equation that indicates the conditions under which the pricing reversal $\underline{p}_I > \overline{p}_I$ occurs just as we did previously. Furthermore, precisely the same qualitative statements about elasticities hold in the breakeven welfare-maximization model as held in the profit-maximization model.

---

[5] See Mohring [20] for another discussion of this problem, which neglects, however, the pricing reversal phenomenon. See also Mowery [21].

[6] These rules are of the same form as those derived in the more general model of optimal departures from marginal cost pricing, as described in Baumol and Bradford [4] and Boiteux [6]. It is no accident that pricing rules of the Baumol and Bradford variety have appeared in our model. By implicitly assuming that we know which of the periods is peak and which is offpeak (as we have in our models), it has been possible to avoid explicit introduction of the capacity constraints, $\underline{X} \leqq X$ and $\overline{X} \leqq X$. This is precisely the trick that is needed to reduce the peak-load pricing problem to a special case of an optimal departures from marginal cost pricing problem.

?

3. The regulated firm

■ **Breakeven regulation.** We now consider a peak-load pricing model with profit maximization as the objective but with a regulatory constraint on the firm's behavior. If the constraint binds the firm in such a way that its total revenue just covers its total cost, including the cost of capital, then no unique solution to the peak-load pricing problem exists. With the profit goal, any pair of peak and offpeak prices consistent with the zero-profit constraint is equally desirable to management.

☐ **Rate-of-return regulation.** If the Averch-Johnson[7] assumption of a fair return s, larger than the market cost of capital r, is made, then a unique pair of peak and offpeak prices can be attained. Model (10) in Table 1 follows Bailey[8] but uses the idea from Bailey[9] that operation off the production frontier is a possible alternative for the firm.[10] We use λ to denote the multiplier associated with the rate-of-return constraint, Z for the number of units of capital wasted, and the subscript G to denote the solution prices for the model.

Equations (11) and (12) give the rules for peak and offpeak prices. Equation (12) is precisely the same as the rule followed by the unregulated profit-maximizing firm [cf., equation (12) and equation (6)]. Thus, in the absence of cross-elasticity effects, (12) says that the offpeak quantity and price are identical to that which an unregulated monopoly would attain; the constraint does not affect the offpeak policy.[11]

The entire effect of regulation is reflected instead in peak price changes and/or in the possibility of operating off the production frontier. Equation (11) gives the peak-period effect. Marginal revenue is set equal to something less than the marginal cost in the peak period. Intuitively, because constrained profits increase with the level of capacity, the firm passes on the benefits of regulation to those users whose increased demand will cause an increase in capacity. *Thus, the peak period price under rate-of-return regulation will be lower than the peak period price in a model of unconstrained profit maximization* [*cf., (11) and (5)*]. *Accordingly,*

---

[7] In [1].

[8] In [3].

[9] In [2].

[10] The idea that the firm might choose to operate off its production possibility frontier in a peak-load pricing model also appears in Riley [25]. See also Zajac [32]. In model (10) the regulator is assumed not to be able to distinguish between $\overline{X}$ and Z and thus permits a fair return s on either type of capital expense. If the regulator is instead assumed to detect and disallow all Z, the conclusions change slightly, as we shall shortly show.

[11] Note, however, that if there are cross-elasticities, the solution to (12) depends on the value of s, since there is a term coupling peak and offpeak demands

$$\overline{X}\frac{\partial \overline{p}}{\partial \underline{X}}.$$

Hence, the act of regulation may be expected to affect offpeak quantity and price in any case where a customer utilizes the service in both peak and offpeak periods.

the likelihood is increased of finding the pricing reversal described in Section 2.

This model also has some interesting implications as to the capacity that is provided.[12] Let us denote the surplus of offpeak revenues over offpeak operating costs as

$$\underline{S} = \underline{p}_G\underline{X}_G - c\underline{X}_G,$$

and let us assume for the moment that the firm operates on the production frontier ($Z_G = 0$). Then the constraint in model (10) can be rewritten

$$\underline{S} + \bar{p}_G\bar{X}_G - c\bar{X}_G - rk\bar{X}_G = (s - r)k\bar{X}_G\,,$$

which can be used to determine the conditions under which the peak-period capacity of the regulated monopoly would exceed or fall short of the socially optimal level of capacity.

This equation tells us that once the surplus from the offpeak period has been determined, the firm portrayed by this model would choose the peak-period quantity (= capacity) and price to bring its *overall* profits in line with the constraint. In choosing its peak-period capacity, the firm would thereby determine its rate base and the level of allowable profits under the constraint. The surplus from the offpeak period (already determined) and the surplus or deficit from the peak period (equal to peak revenues minus the sum of peak operating costs and total capacity costs) must together sum to the total allowable profits under the constraint.

More specifically, if the offpeak surplus is very large, the firm would lower its peak price below $c + rk$, and thus run a deficit on its peak period operations. It would do this in order to absorb its offpeak surplus and at the same time to expand its capacity and thus expand its allowable profits. The firm would continue to lower its price and expand its capacity until the allowable profits on the expanded rate base equal the offpeak surplus less the peak period deficit. The rate-of-return constraint has given the firm an incentive to absorb its extra profits through expanded peak capacity rather than changing its offpeak price to reduce profits.[13]

If, in contrast, there is no offpeak surplus, $\underline{S} = 0$, then the firm sets peak price, $\bar{p}_G = c + sk$. Thus, peak price lies between the monopoly and the welfare levels with capacity having been increased toward, but not beyond, that which is socially desirable. If there is some offpeak surplus, but it is comparatively small, the firm will lower peak price from $c + sk$, still making a surplus on its peak operations, until the sum of the offpeak and peak surplus comes into equality with the allowable profits.

---

[12] See Wellisz [30] for the earliest treatment of this matter.

[13] In the more general case where each class of customers uses the facilities during both peak and offpeak periods, the regulatory constraint will almost certainly affect offpeak prices to at least some extent (see, for example, Bailey [3] and Currie [11]). Offpeak price will also be affected if Klevorick's [17] "labor-requirements function" is used instead of the constant expense, c, in the cost term. In this case, the substitution of capital for labor in production will serve to lower the marginal labor cost, and hence will reduce offpeak price as well as peak price as the firm moves out along the Averch-Johnson [1] expansion path. See also Waverman [29].

These last results expand considerably on one of Bailey's principal conclusions. Bailey[14] shows that the peak-period capacity may exceed the welfare-maximization level if the cross-elasticity is not zero. Here, we see a more general result: *Peak-period capacity exceeds the welfare-maximization level whenever the surplus from the offpeak period is large enough to require deficits in the peak period.*[15] *In the other case, in which there is some surplus in the peak period, the regulated firm will provide a larger capacity than would an unregulated monopolist,* $\overline{X}_M < \overline{X}_G$, *but below that which welfare maximization produces,* $\overline{X}_G < \overline{X}_W$. The crucial variables determining whether or not the peak capacity exceeds the welfare maximizing level are the size of the offpeak surplus and the *margin* between s and r. As the margin between s and r is reduced, the constrained level of profits declines, and it becomes more likely that the offpeak surplus will have to be absorbed by an expansion of peak-period capacity and consequent peak-period deficits. Note, though, that if the margin between s and r narrows because s is lowered, peak-period capacity will expand; whereas, if the margin narrows because r rises, the peak-period capacity will not change but the welfare-maximizing capacity level will decline. These last results follow from the constraint in model (10), in which s, but not r, is a determinant of the peak-period capacity.[16]

We now examine the possibility of $Z > 0$, i.e., the possibility that the firm is operating off its production frontier by employing excess capital. If $Z > 0$, then the Kuhn-Tucker condition on the variable Z requires that $\lambda = \frac{r}{s}$.[17] Substituting into (11), we can see that this happens only when

$$\overline{p} = \frac{c}{1 - \frac{1}{\overline{e}}}.$$

The common sense of this equation is straightforward. The firm wishes to expand its capacity to absorb its offpeak surplus. If it is worthwhile to expand capacity at the point where the marginal revenue equals marginal operating cost, then the best way of expanding capital further is to add capital that is not used productively. Such an addition of unproductive capital does not necessitate any price drop, whereas adding productive capacity means that the price and marginal revenue on peak operation have to drop and the firm's marginal expenditure on operating costs exceeds the additional revenue it brings in.[18] For the rest of this section, we shall suppose that $\overline{MR}_G > c$ at the optimal solution, so that operation off the production frontier is not a profitable alternative.

---

[14] In [3], p. 676.

[15] Kafoglis [16] also recognizes the possibility that expansions in capacity beyond the welfare-maximizing level can occur in regulatory models when there is price discrimination between classes of customers.

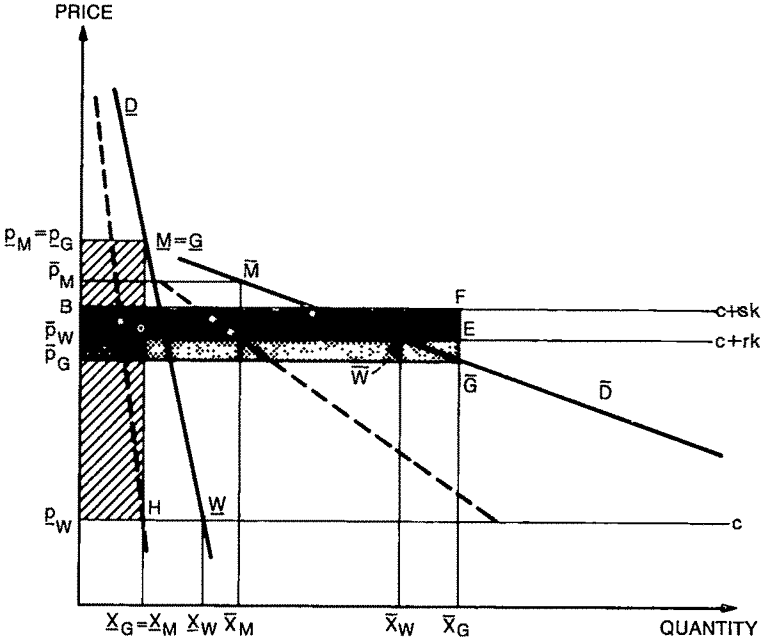
[16] A similar sort of result is derived in Bailey [2], Chapters 6 and 8, in which it is shown that in an Averch-Johnson [1] model of rate-of-return regulation the optimal capital stock increases as the fair return s is lowered toward r, but does not change if the regulator holds s fixed and the cost of capital r fluctuates with changes in market conditions. See also Baumol and Klevorick [5].

[17] The Kuhn-Tucker conditions on Z are that $k(-r + \lambda s) \leqq 0$, and that $Zk(-r + \lambda s) = 0$.

The results under rate-of-return regulation may be seen graphically in Figure 3. As before, the profit-maximizer's solution is denoted $\overline{M}$ and $\underline{M}$, and the welfare solution is $\overline{W}$ and $\underline{W}$. The horizontal line $c + sk$ is constructed so that the difference between it and the $c + rk$ line is precisely the permitted return $(s - r)k$ per unit of capacity.

FIGURE 3

PRICING UNDER RATE-OF-RETURN REGULATION



For the regulated firm, constrained profits at any given peak period output $\overline{X}_G$ (= capacity) would be equal to $(s - r)k\overline{X}_G$, or the solid rectangle $\overline{p}_W BFE$. If, as in the diagram, the surplus from the offpeak period, the striped rectangle, $\underline{p}_G GH\underline{p}_W$, exceeds $\overline{p}_W BFE$, the firm would set a price $\overline{p}_G$ for the peak period that was below $c + rk$ until the deficits, the dotted rectangle $\overline{p}_W EG\overline{p}_G$, were such that $\underline{p}_G GH\underline{p}_W - \overline{p}_W EG\overline{p}_G = \overline{p}_W BFE$. This last condition can also be restated as $\underline{p}_G GH\underline{p}_W = \overline{p}_W BFE + \overline{p}_W EG\overline{p}_G$ or $\underline{p}_G GH\underline{p}_W = BFG\overline{p}_G$, i.e., the striped rectangle is equal to the sum of the solid and the dotted rectangles in Figure 3. In effect, then, the regulated firm finds that point on its peak demand curve $\overline{G}$ at which the product of the *margin* between $c + sk$ and $\overline{p}_G$, times the $\overline{X}_G$ that is forthcoming at that $\overline{p}_G$, is just equal to the offpeak surplus. This result has a simple interpretation in common-sense terms: $c + sk$ represents the maximum that the firm could charge for peak sales if it had no offpeak surplus. But if the firm has an offpeak surplus, it cannot charge the full $c + sk$ for its peak operations. Hence, it drops its peak price until the difference between $c + sk$ and its peak price, times the peak quantity, can just absorb the offpeak surplus.

---

[18] Of course, if the regulator detects and disallows unproductive capital, such expenditures only subtract from profits and do not add to the rate base. In this case, adding productive capital would always be preferable.

As is clear from the diagram, if the margin $(s - r)k$ (the distance FE) is smaller than the distance $F\overline{G}$, the firm will be running a deficit in the peak period and the regulated firm will be providing peak capacity that is larger than that which maximizes welfare. The opposite will be true if the distance FE is larger than the distance $F\overline{G}$. Note, though, as was argued before, if r moves relative to s, the peak capacity offered does not change (though the welfare-maximizing level changes). Thus, in the diagram, if r increases so that the margin $(s - r)k$ is diminished, the peak output $\overline{X}_G$ does not change. This is true because $\overline{X}_G$ is still the only point at which $BF\overline{Gp}_G = \underline{p}_G\underline{G}H\underline{p}_W$. All that is implied is that the firm is running larger losses on its unchanged peak output and its overall profits are reduced. If, on the other hand, s declines, $BF\overline{Gp}_G$ is no longer equal to $\underline{p}_G\underline{G}H\underline{p}_W$, and $\overline{X}_G$ has to increase.

4. Two-part tariffs

■ **Perfect two-part tariffs.** The preceding sections have described models in which only one charge is levied on peak users and one on offpeak users. In practice, however, firms such as public utilities usually have a set of multiple prices or block tariffs with prices falling in steps as greater volumes are used. To reflect such a system of prices, we now introduce a model of peak-load pricing in which a customer charge as well as a usage charge is levied.

To avoid complicating the model, we suppose that the firm can charge an entrance fee or customer charge to each class of customers, as well as a peak and an offpeak usage price. Furthermore, we assume that all customers in a particular customer class have identical demand curves.[19]

If the firm has complete freedom in setting the customer fees, $\overline{E}$ and $\underline{E}$, and the usage fees, $\overline{p}$ and $\underline{p}$, the model becomes

$$\begin{aligned} \underset{\overline{E},\underline{E},\overline{X},\underline{X},Z}{\text{Maximize}} \quad \pi^* = {} & \overline{E} + \underline{E} + \overline{p}(\overline{X})\overline{X} + \underline{p}(\underline{X})\underline{X} - c(\overline{X} + \underline{X}) \\ & -rk(\overline{X} + Z) \end{aligned}$$

$$\begin{aligned} \text{subject to } \overline{E} &\leqq \int_0^{\overline{X}} \overline{p}(\overline{X})\,d\overline{X} - \overline{p}(\overline{X})\overline{X} \qquad (13) \\ \underline{E} &\leqq \int_0^{\underline{X}} \underline{p}(\underline{X})\,d\underline{X} - \underline{p}(\underline{X})\underline{X} \\ \pi^* &\leqq (s - r)k(\overline{X} + Z). \end{aligned}$$

The constraints on the entrance fees assert that these cannot exceed the accumulated consumers' surplus above the marginal usage charge to the particular customers. For, otherwise, the customer would purchase none of this particular product or service. If $\overline{\alpha}$

---

[19] This is the same sort of assumption as is made in MacAvoy and Noll [19]. The cases where customers have different demand curves and/or where entrance fees can be tailored to individual customers have been described by Oi [22]; however, Oi does not include peak and offpeak pricing considerations.

and $\underline{\alpha}$ denote the Lagrange multipliers for the entrance fee constraints, the Lagrangian for model (13) becomes

$$\begin{aligned}\phi(\bar{E},\underline{E},\bar{X},\underline{X},Z,\lambda,\bar{\alpha},\underline{\alpha}) = {} & (1-\lambda)[\bar{E}+\underline{E}+\bar{p}(\bar{X})\bar{X} \\ & +\underline{p}(\underline{X})\underline{X}-c(\bar{X}+\underline{X})-rk(\bar{X}+Z)] \\ & +\lambda(s-r)k(\bar{X}+Z) \\ & +\bar{\alpha}\left[\int_0^{\bar{X}}\bar{p}(\bar{X})d\bar{X}-\bar{p}(\bar{X})\bar{X}-\bar{E}\right] \\ & +\underline{\alpha}\left[\int_0^{\underline{X}}\underline{p}(\underline{X})d\underline{X}-\underline{p}(\underline{X})\underline{X}-\underline{E}\right]. \qquad (14)\end{aligned}$$

The Kuhn-Tucker conditions are

$$\phi_{\bar{E}}:\quad 1-\lambda\leqq\bar{\alpha},\ \bar{E}(1-\lambda-\bar{\alpha})=0 \qquad (15)$$

$$\phi_{\underline{E}}:\quad 1-\lambda\leqq\underline{\alpha},\ \underline{E}(1-\lambda-\underline{\alpha})=0 \qquad (16)$$

$$\begin{aligned}\phi_{\bar{X}}:\quad & (1-\lambda)\left[\bar{p}(\bar{X})+\bar{X}\frac{d\bar{p}}{d\bar{X}}-c-rk\right] \\ & +\lambda(s-r)k-\bar{\alpha}\left(\bar{X}\frac{d\bar{p}}{d\bar{X}}\right)=0 \qquad (17)\end{aligned}$$

$$\phi_{\underline{X}}:\quad (1-\lambda)\left[\underline{p}(\underline{X})+\underline{X}\frac{d\underline{p}}{d\underline{X}}-c\right]-\underline{\alpha}\left[\underline{X}\frac{d\underline{p}}{d\underline{X}}\right]=0 \qquad (18)$$

$$\phi_{Z}:\quad -rk+\lambda sk\leqq 0,\ Zk(-r+\lambda s)=0 \qquad (19)$$

$$\begin{aligned}\phi_{\bar{\alpha}}:\quad & \bar{E}\leqq\int_0^{\bar{X}}p(\bar{X})d\bar{X}-\bar{p}(\bar{X})\bar{X}, \\ & \bar{\alpha}\left(\bar{E}-\int_0^{\bar{X}}\bar{p}(\bar{X})d\bar{X}-\bar{p}(\bar{X})\bar{X}\right)=0 \qquad (20)\end{aligned}$$

$$\begin{aligned}\phi_{\underline{\alpha}}:\quad & \underline{E}\leqq\int_0^{\underline{X}}\underline{p}(\underline{X})d\underline{X}-\underline{p}(\underline{X})\underline{X}, \\ & \underline{\alpha}\left(\underline{E}-\int_0^{\underline{X}}\underline{p}(\underline{X})d\underline{X}-\underline{p}(\underline{X})\underline{X}\right)=0 \qquad (21)\end{aligned}$$

$$\phi_{\lambda}:\quad \pi\leqq(s-r)k(\bar{X}+Z),\ \lambda[\pi-(s-r)k(\bar{X}+Z)]=0. \qquad (22)$$

A number of interesting results emerge from these conditions. First, in the absence of regulation, $\lambda=0$ and, from (15) and (16), $\bar{\alpha}=\underline{\alpha}=1$. Then, by (20) and (21), *the firm sets the customer charge for each class of customers so as to equal precisely the entire area above the usage price and below the demand curve.* Substituting $\bar{\alpha}=\underline{\alpha}=1$ into (17) and (18), and simplifying, we find

$$\bar{p}_T=c+rk \qquad (23)$$

$$\underline{p}_T=c, \qquad (24)$$

so that *under a two-part tariff welfare pricing is preserved: the usage prices to the peak and offpeak customers just exactly compensate at the margin for their respective additional costs.* Stated differently, the elasticity terms that appeared in equations (5) and (6) have disappeared, and the welfare-maximization results of

Section 2 have reemerged. This result is intuitively obvious once it is recognized that when $\overline{E}$ and $\underline{E}$ are set at their upper bounds, the "profits" in the objective function of model (13) are precisely the social welfare returns from model (1). With model (13) restated in the form

$$\underset{\overline{E},\underline{E},\overline{X},\underline{X},Z}{\text{Maximize}} \quad W = \overline{E} + \underline{E} + \pi$$

$$\text{Subject to} \quad W \leqq (s-r)(\overline{X}+Z),$$

it is clear that the elasticities of demand will play no role in the analysis.

The welfare result holds also if K' replaces rk as the marginal cost of capacity. Thus, in decreasing average cost industries, a two-part tariff enables the firm to be financially viable without having to depart (optimally or otherwise) from marginal cost pricing. Welfare pricing ($\overline{p}_W$ and $\underline{p}_W$ in Figure 2) is the outcome, and yet the firm is not supported by taxation out of the public treasury.[20] *The two-part tariff is indeed "perfect" both because all of the consumers' surplus is extracted and because the investment decision, consumption decision, and resource allocation decision are all made at the correct margin. However, the income distribution is now very different, since the utility has captured all of the social surplus*[21] (the peak surplus $A\overline{Wp}_W$ in Figure 2 plus the offpeak surplus $E\underline{Wp}_W$ less the capacity deficit $B\overline{Wp}_W$.)

☐ **Regulation of rate of return but not of price structure.** One way to redistribute the income is to impose a rate-of-return constraint. Such a constraint has been included in model (13). The model is thus one in which the regulator permits the firm to set its own price structure and insists only that the earned rate of return be at the level deemed fair. When regulation is effective, $0 < \lambda < 1$ from (19). From (15) and (16), this means that $\overline{\alpha}$ and $\underline{\alpha}$ are both nonzero, so that the equalities must hold in the entrance fee constraints (20) and (21): as before, the customer charges are set so as to absorb the entire surplus.

If we solve for the offpeak usage price by substituting $\underline{\alpha} = 1 - \lambda$ into (18), we achieve the same result as (24): the offpeak user is charged his marginal running cost. To solve for the peak price, substitute $\overline{\alpha} = 1 - \lambda$ into (17) to obtain

$$\overline{p}_R = c + rk - \frac{\lambda}{1-\lambda}(s-r)k. \tag{25}$$

Hence, the price to the peak user is set below the sum of marginal operating and capacity costs and *capacity will always be in excess of the welfare-maximizing level. In essence, the regulated utility*

---

[20] This point was made by Coase [10], p. 118, and by Buchanan [8], p. 470; also, see Oi [22] and Buchanan [9]. Wallace Oates pointed out to us that for those in the field of public finance, welfare maximization implies a "regulatory" policy which actually leads to the welfare-maximizing outcome, which of course is precisely what happens in the two-part tariff situation.

[21] For an interesting treatment of income distribution in this context, see Feldstein [14].

absorbs the surplus from the customer charges entirely by expanding capacity.

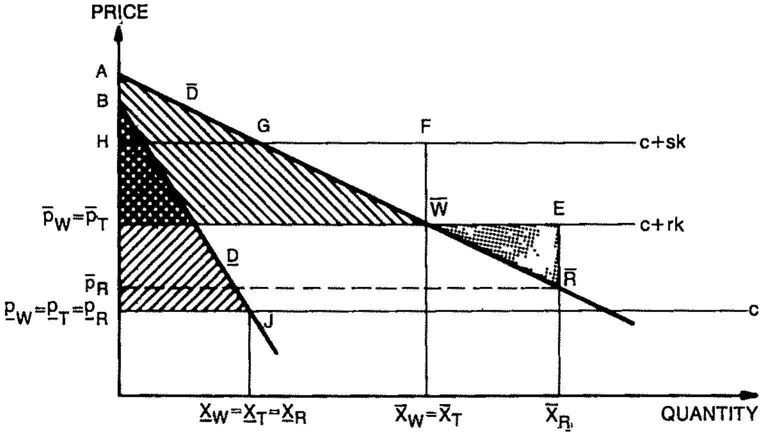
This expansion of capacity will be productive as long as $\bar{p}_R > c$. To see this, note that if waste is positive ($Z > 0$), then $\lambda = r/s$, so that from (17), price must have been driven down to marginal operating cost. As long as price is above this marginal cost, the firm prefers to add productive capacity, for by doing so it is simultaneously earning a larger surplus from the peak customer charge and is adding to the rate base. If price is below marginal running costs, then a deficit is incurred on marginal operation and it is cheaper, as long as the regulator permits Z to enter the rate base, for the firm to add unused capital.

These results indicate that, *if the regulator does not distinguish productive from nonproductive capital, the pricing reversal discussed in previous sections cannot occur.* The price to offpeak users is set equal to marginal operating cost, and the price to the peak user is set above or equal to this level.

Figure 4 displays the equilibrium graphically. The regulated firm sets the same offpeak customer charge (equal to the area of the triangle $BJ\underline{p}_R$) and the same price, $\underline{p}_R$ as the unregulated firm. The offpeak surplus, along with that of the peak user (given by the triangle $A\bar{W}\bar{p}_W$), yields profits in excess of those allowed at the welfare-maximizing level of capacity, i.e., $A\bar{W}\bar{p}_W + BJ\underline{p}_R > HF\bar{W}\bar{p}_W$. Hence, the utility lowers peak price and expands capacity until at capacity $\bar{X}_R$ the total profit has been reduced (by the dotted triangle $\bar{W}E\bar{R}$) sufficiently to bring about equality in the regulatory constraint. The equilibrium price to the peak user, $\bar{p}_R$, is clearly below the welfare-maximizing level.

FIGURE 4

TWO-PART TARIFF AND RATE-OF-RETURN REGULATION



☐ **Regulation of price structure.** In practice, the fixed charges levied by utilities appear to fall far short of the full consumer's surpluses. It is likely that there are formal or informal limits imposed by regulators on the fixed charges that can be levied.[22] Let us suppose that the customer fees are set so as to take into account the large fixed capital costs in the industries, but that they are limited to levels below those at which all the surplus is absorbed. The interesting feature of the analysis when we construct the model in this way is that we revert to conclusions that closely resemble those in Section 3. The elasticities of the different classes of customers are again evident, and the pricing reversal possibility reappears.

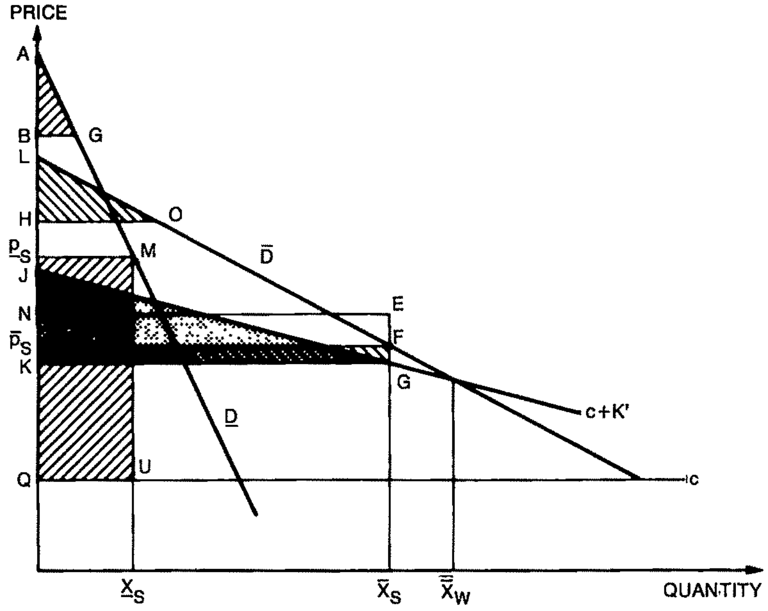
---

[22] These limits may arise in part because customers have different de-

Figure 5 illustrates the situation. The customer charge to the offpeak user is assumed to be limited by the regulator to area *AGB.*

FIGURE 5

PRICING WITH LIMITED CUSTOMER CHARGE



A profit-maximizing firm does not set usage price equal to marginal operating cost for this offpeak period, but instead sets marginal revenue equal to cost, for by doing so, the offpeak surplus $\underline{p}_S MUQ$ can be gained in addition to the offpeak customer charge, *AGB*. It is for this reason that the elasticity effects show up in the first-order conditions.[23]

Similarly, the firm will treat the customer fee of the peak users, triangle *LOH* in Figure 5, as an endowment of profit, and will

---

mand curves. In order to make the service available to everyone, the entrance fee cannot exceed the level which the customer with the smallest consumer's surplus is willing to pay. See Oi [22].

[23] More specifically, if $\bar{E}$ and $\underline{E}$ are limited to amounts unrelated to the elasticities of demands, then the model becomes

$$\underset{\bar{X},\, \underline{X},\, Z}{\text{Maximize}} \quad \pi + \text{constant}$$

$$\text{Subject to } \pi + \text{constant} \leqq (s - r)k(\bar{X} + Z).$$

Thus, it is not surprising that these elasticities will reappear in the first-order conditions determining the prices for peak and offpeak service. Elasticity effects would only be modified if the regulator permitted the customer fee to be in excess of a monopoly surplus triangle such as $AM\underline{p}_S$ in Figure 5.

proceed to adjust the usage price to the peak user until the regulatory constraint is satisfied. If the fair return is such that the profit permitted by the constraint is given by rectangle *NEGK*, then this rectangle must equal the sum of the surplus from the customer charges $AGB + LOH$, the surplus from the offpeak usage charge, $\underline{p}_S MUQ$, the deficit or surplus from the peak period (the surplus $\bar{p}_S FGK$), less the deficit *JGK* incurred by the increasing returns to capacity, that is,

$$NEGK = AGB + LOH + \underline{p}_S MUQ + \bar{p}_S FGK - JGK.$$

5. Summary

■ We have demonstrated in this paper that it is theoretically possible for offpeak prices to be higher than peak prices for an unconstrained profit-maximizing firm, for a welfare-maximizing firm with increasing returns to scale, for a profit-maximizing firm subject to a regulatory constraint on its rate of return, and for a firm that can levy a two-part tariff but where there is a constraint on the size of its customer charge. In the first two instances, the pricing reversal comes about because of the inverse elasticity rule, familiar from the optimal departures from marginal cost pricing literature. In the latter cases, the reversal comes about because the rate-of-return constraint encourages price reductions to peak-period users rather than to offpeak users. As a result, the pricing reversal becomes ever more likely to occur as the constraint is tightened. Furthermore, as the constraint is tightened, it becomes more likely that there can be an expansion of capacity beyond the welfare-maximizing level.

In the electricity example, these last results imply that for customer charges of approximately the same size, state regulatory agencies that have tighter rate-of-return requirements may encourage lower usage prices to peak business users while the prices charged to offpeak residential users remain substantially unchanged. An additional implication is that effectively regulated privately owned electric utilities might grant larger price reductions for business users than those instituted by government-owned electrics.

Either of these views suggests a possible method of empirically testing whether actual behavior in the electric utility industry agrees with that predicted by the theoretical models of peak-load pricing reversals. Recent evidence confirms that such agreement may be found. Peltzman, in contrasting the behavior of regulated and government-owned electrics finds that "residential output is about ⅓ greater . . . [while] industrial output is twice as large in privately owned as in government-owned utilities . . . [so that] the magnitude of the output effect of ownership is, in fact, smallest for residential customers."[24] The relatively larger price reductions to the business users in the regulated electrics is just what our theory would predict.

---

[24] Peltzman [23], p. 138. Strictly speaking, as Peltzman pointed out to us in correspondence, his "result was based upon more finely developed price discrimination by private as opposed to public utilities, rather than upon any difference in average residential and industrial rates." However, he goes on to state "the thrust of your argument seems to me to imply that regulated utilities will be motivated to seek special rates for the more elastic demands within as well as between customer groups."

References

1. Averch, H. and Johnson, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review,* Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. Bailey, E. *Economic Theory of Regulatory Constraint.* Lexington, Mass.: Heath-Lexington, 1973.
3. ———. "Peak-Load Pricing under Regulatory Constraint." *Journal of Political Economy,* Vol. 80, No. 4 (July/August 1972), pp. 662–679.
4. Baumol, W. J. and Bradford, D. F. "Optimal Departures from Marginal Cost Pricing." *The American Economic Review,* Vol. 60, No. 3 (June 1970), pp. 265–283.
5. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
6. Boiteux, M. "On the Management of Public Monopolies Subject to Budgetary Constraints." *Journal of Economic Theory,* Vol. 3, No. 3 (September 1971), pp. 219–240.
7. ———. "Peak-Load Pricing." *Journal of Business,* Vol. 33, No. 2 (April 1960), pp. 157–179.
8. Buchanan, J. M. "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics,* Vol. 80, No. 3 (August 1966), pp. 472–480.
9. ———. "The Theory of Monopolistic Quantity Discounts." *Review of Economic Studies,* Vol. 20 (1952–1953), pp. 199–208.
10. Coase, R. H. "The Theory of Public Utility Pricing and Its Application." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 1 (Spring 1970), pp. 113–128.
11. Currie, K. A. "Peak Load Pricing under Regulatory Constraint: Comment." Unpublished manuscript, 1973.
12. Davidson, R. K. *Price Discrimination in Selling Gas and Electricity.* Baltimore: Johns Hopkins Press, 1955.
13. *Electric Power and Government Policy* (New York: The Twentieth Century Fund, 1948).
14. Feldstein, M. S. "Equity and Efficiency in Public Sector Pricing: The Optimal Two-Part Tariff." *Quarterly Journal of Economics,* Vol. 86, No. 2 (May 1972), pp. 175–187.
15. Hirshleifer, J. "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics,* Vol. 72, No. 3 (August 1958), pp. 451–462.
16. Kafoglis, M. Z. "Comment," in Harry M. Trebing, ed., *Essays on Public Utility Pricing and Regulation,* East Lansing: Michigan State University, 1971.
17. Klevorick, A. K. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 122–153.
18. Littlechild, S. C. "Peak Load Pricing of Telephone Calls." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 191–210.
19. MacAvoy, P. W. and Noll, R. "Relative Prices on Regulated Transactions of the Natural Gas Pipelines." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 212–234.
20. Mohring, H. "The Peak Load Problem with Increasing Returns and Pricing Constraints." *The American Economic Review,* Vol. 60, No. 4 (September 1970), pp. 693–705.
21. Mowery, V. O. "Optimal Pricing Theories for Interrelated Monopoly Outputs." Bell Laboratories Technical Memorandum, November 1970.

22. Oi, W. Y. "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly." *Quarterly Journal of Economics,* Vol. 85, No. 1 (February 1971), pp. 77–96.
23. Peltzman, S. "Pricing in Public and Private Enterprises: Electric Utilities in the United States." *Journal of Law and Economics,* Vol. 14, No. 1 (April 1971), pp. 109–147.
24. Pressman, I. "Peak Load Pricing." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 304–326.
25. Riley, J. G. "Peak Load Pricing under Regulatory Constraint: Comment." Unpublished manuscript, 1972.
26. Steiner, P. O. "Peak Load Pricing Revisited," in Harry M. Trebing, ed., *Essays on Public Utility Pricing and Regulation,* East Lansing: Michigan State University, 1971.
27. ———. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics,* Vol. 71, No. 4 (November 1957), pp. 585–610.
28. Turvey, R. "Peak-Load Pricing." *Journal of Political Economy,* Vol. 76, No. 1 (January/February 1968), pp. 107–113.
29. Waverman, L. "Peak Load Pricing under Regulatory Constraint—A Proof of Inefficiency." Unpublished manuscript, 1973.
30. Wellisz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy,* Vol. 71, No. 1 (February 1963), pp. 30–43.
31. Williamson, O. E. "Peak Load Pricing and Optimal Capacity under Indivisibility Constraints." *The American Economic Review,* Vol. 56, No. 4 (September 1966), pp. 810–827.
32. Zajac, E. E. "Note on 'Gold Plating' or 'Rate Base Padding.'" *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 311–315.

The effect of regulation changes on insider trading

Jeffrey F. Jaffe

The Wharton School
University of Pennsylvania

Though lawyers and economists have speculated about the consequences of effective insider regulation, empirical research has yet to establish whether the regulation of insiders is even effective. Following the Securities and Exchange Act of 1933–1934, the most significant changes in insider regulation have occurred through a few important decisions in case law. This study examines the changes in the volume and profitability of insider trading after each of three important legal decisions concerning insiders. No significant change in the properties of insider trading was observed following any of these three events. In addition, there appeared to be no combined effect from the three decisions. Thus it is concluded that the null hypothesis that changes in regulation had no effect on the trading of insiders cannot be rejected.

1. Introduction

■ The effect of government regulation on the trading of insiders is a matter of great concern in legal circles, as the proliferation of law journal articles and court cases on the subject will testify. The legal literature has been concerned primarily with the presumed injustice of knowledgeable insiders' benefiting at the expense of other investors.

Economists, by contrast, typically have been more involved with the effect of insiders on capital market efficiency[1]—and there is much disagreement. On the one hand, Friend states that one purpose of security regulation is "to prevent insiders' profiting thereby at the expense of shareholders."[2] He concludes that "I

Jeffrey E. Jaffe received a B.A. in economics from the University of Chicago in 1968. He received the M.B.A. degree in 1971 and the Ph.D. degree in 1972 from the Graduate School of Business at the University of Chicago. He is presently Assistant Professor of Finance at the Wharton School of the University of Pennsylvania, where his current interest is in the value of information in speculative markets, optimal trading rules in a market with a specialist, the theory of finance under uncertainty and personal taxes, and the performance of new issues.

Much of this paper is drawn from my Ph.D. thesis, "Special Information, Security Regulation and Insider Trading." I would like to thank my dissertation committee, Fischer Black, Eugene Fama and Merton Miller (chairman) for their advice and encouragement throughout all the stages of my thesis. The insightful comments of G. Mandelker and R. Ibbotson have also been quite helpful. I would further like to thank the finance workshops at the University of Chicago, Stanford University, Columbia University, Indiana University and the University of Pennsylvania.

[1] For an excellent discussion of the differing viewpoints of lawyers and economists, see Manne [38], p. 3.

[2] See Friend [23], p. 118.

have never thought there was any equity or economic justification for revoking that part [of the Securities and Exchange Act] which deals with disclosure of insider trading."[3] On the other hand, Demsetz[4] argues that secrecy is not necessarily an evil, for it prevents "the rapid communalization of rewards"[5] from information. Demsetz also views the costs of policing insider trading as substantial. Manne[6] adds that regulation may actually increase secrecy by lengthening the time between the production of information and its release to the market.

Though lawyers and economists have speculated about the consequences of effective insider regulation, empirical research has yet to establish whether the regulation of insiders is even effective. The techniques of this study permit judgments to be made on the effectiveness of regulation.

The results of previous studies suggest that insiders possess special information.[7] The present study compares the extent of their use of that special information in various time periods with the degree of regulation in effect during these periods. After the Securities and Exchange Act of 1933–1934, all important changes in insider trading regulation resulted from case law.[8] Three of the most significant regulatory events concerning insider trading are:[9]

1. the Cady, Roberts decision (November 8, 1961),
2. the Texas Gulf Sulphur indictment (April 19, 1965), and
3. the Texas Gulf Sulphur decision (August 19, 1966).

These three important cases signaled changes in the attitude of the courts and law enforcement agencies towards insider trading. Therefore, we shall compare the behavior of insider trading immediately before and after each of these three regulatory changes. This technique should reduce the effect of other economic influences, such as changing market conditions, which have plagued studies comparing periods separated by many years. The combined effect of the three decisions is also measured.

The format of the paper is as follows. In Section 2, a method for estimating the performance of insider trading is presented. In Section 3 this method is used to measure the performance of insiders in selected samples. The effectiveness of particular decisions in case law is investigated in Section 4. We conclude with a discussion of the implications of this paper for public policy and future research. A history of the regulation of insiders is presented in the appendices.

---

[3] *Ibid.*, p. 118.

[4] See Demsetz [13], pp. 1–23.

[5] *Ibid.*, p. 11.

[6] Manne [38], p. 88.

[7] Evidence on the profitability of insider trades is presented in this paper. The reader is also referred to Driscoll [14], Glass [24], Lorie and Neiderhoffer [34], and Rogoff [42].

[8] A short history of security regulation is presented in the Appendix.

[9] In the matter of Cady, Roberts and Company, SEC Release No. 6668, CCH Federal Securities Law Reporter par 76,803 (1961); SEC v. Texas Gulf Sulphur Company, 401 F. Supp. 262 (S.D.N.Y. 1966). The importance of these cases is discussed in the Appendix.

Method for calculating the "residual" of a security. In order to estimate the profitability of insider trades, this study examines the performance of a security subsequent to specific types of insider trades in that security, which we call insider trading events.[10] We define the performance of a security to be that part of its return not explained by price movement in the stock market as a whole. The performance of a security, when measured in this way, is commonly referred to as the security's residual.[11] The method used in this study to calculate the residual is explained below. Next, statistical measurements and tests which are based on the concept of the residual are discussed.

Sharpe[12] and Lintner[13] have developed a model of equilibrium pricing of risky capital assets. This model postulates a linear relationship between the expected return on a security and the covariance of the security's return with the return on a portfolio composed of all securities in the market (commonly called the market portfolio) and can be expressed as follows:

$$E(\tilde{R}_{j,t}) = R_{f,t} + [E(\tilde{R}_{m,t}) - R_{f,t}]\,\beta_{j,t},$$

where:

$R_{f,t}$ = Rate of return on a riskless asset during period t.

$\tilde{R}_{j,t}$ = Rate of return on security j during period t.

$\tilde{R}_{m,t}$ = Rate of return on the portfolio of all assets during period t.

$\beta_{j,t}$ = $\mathrm{Cov}(\tilde{R}_j,\tilde{R}_m)/\sigma^2(\tilde{R}_m)\cdot\beta_{j,t}$ is the risk of security j measured relative to the risk of the market portfolio. $\beta_{j,t}$ is proportional to the contribution of the jth security to the risk of the market portfolio.

As the Sharpe-Lintner model may be criticized for unrealistically assuming that riskless borrowing and lending opportunities exist for all individuals, Black[14] has derived the following equation for the case where no riskless asset exists:

$$E(\tilde{R}_{j,t}) = E(\tilde{R}_{z,t}) + [E(\tilde{R}_{m,t}) - E(\tilde{R}_{z,t})]\,\beta_{j,t},$$

where:

$E(\tilde{R}_{z,t})$ = Expected rate of return on a portfolio whose return has zero covariance with the return on the market.

There are many complex problems in empirically testing either of these models and they will not be discussed here.[15] Rather, we note that the following is a specification of the stochastic process generating period-by-period returns that is consistent with either of the above models and that has empirically been found to provide a good representation of actual returns:[16]

---

[10] Three types of insider trading events are examined: (1) a transaction of an insider; (2) a large transaction of an insider, and (3) a month in which many insiders of a company transact. Precise specifications of these events are presented later.

[11] The concept of a residual is due to Fama *et al.* [17]. Other studies using this concept are Ball [1], Ball and Brown [2], Ibbotson [26], Mandelker [36], and Scholes [43].

[12] See Sharpe [44].

[13] See Lintner [33].

[14] See Black [4].

[15] The reader is referred to the excellent empirical studies of Black, Jensen and Scholes [5] and Fama and MacBeth [19].

$$\tilde{R}_{jt} = \tilde{\gamma}_{ot} + \tilde{\gamma}_{1t}\beta_j + \tilde{U}_{jt}, \tag{1}$$

where

$\tilde{\gamma}_{ot}, \tilde{\gamma}_{1t}$ = Market determined parameters showing the *ex post* relationship between risk and return at time t,

$\tilde{U}_{jt}$ = The disturbance of the jth security at time t,

β_j = The ratio of the covariance between $\tilde{R}_{jt}$ and $\tilde{R}_{mt}$ to the variance of $\tilde{R}_{mt}$. Note that in equation (1), β_j is assumed to remain constant across time.[17]

Equation (1) indicates that the return on security j in period t is a function of the disturbance, $\tilde{U}_{jt}$, which is specific to an individual security, as well as the market-wide variables, $\tilde{\gamma}_{ot}$, and $\tilde{\gamma}_{1t}$. This disturbance is commonly considered to be the abnormal performance of a security, since the effects of $\tilde{\gamma}_{ot}$, $\tilde{\gamma}_{1t}$, and β_j are netted out. Given the value of $\tilde{R}_{jt}$ and estimates for β_j, $\tilde{\gamma}_{ot}$, and $\tilde{\gamma}_{1t}$, we can use equation (1) to calculate, $\hat{U}_{jt}$, an estimate of the disturbance of security j in period t.[18] This estimate serves as the "residual" of the security in our study.

---

[16] See Black, Jensen, and Scholes [5] and Fama and MacBeth [19].

[17] Both Black, Jensen, and Scholes and Fama and MacBeth assume that β_j changes over long periods of time, and therefore they continually update their estimates of β_j. However, studies examining performance over a relatively short period of time typically assume a constant β_j. For example, see Fama *et al.* [17], Ball and Brown [2], Scholes [43], and Mandelker [36]. In the present study, we examine the performance of a given security over eight consecutive months. The β_j's can reasonably be assumed to be constant over this interval.

[18] For this study, the return on security j in month t is defined as:

$\tilde{R}_{jt} = P_{jt} + d_{jt}/P^*_{j,\,t-1}$,

where

P_{jt} = The market price for a share of security j at the end of month t;
d_{jt} = The cash dividends paid on security j in month t;
P^*_{jt} = The price of a share of security j at the end of month t, *after* an adjustment for capital changes in month $t + 1$. See Fisher [21] for a complete discussion.

The return data for this study were obtained from the Stock Price tape of the Center for Research in Security Prices (CRSP) at the University of Chicago. This tape contains return data for each security listed on the New York Stock Exchange during the period from January 1926 to June 1968.

The measure of risk for security j, β_j, is obtained from the following regression, using the 50 months of data prior to the event:

$\tilde{R}_{jt} = \alpha_j + \beta_j \tilde{R}mt + \tilde{E}_{jt}$,

where

$\tilde{R}mt$ = The return on the portfolio of all assets in the market during month t. This study used the Fisher Index (see [22]) for the market portfolio.

Discussion of this regression and tests of its specification can be found

☐ **Statistical measurements.** Based on information gathered from the *Official Summary of Insider Trading,*[19] samples are chosen consisting of securities undergoing a specific insider trading event. (Samples of events are listed in footnote 10.) To estimate the gains from insider trading, residuals subsequent to events are examined. Methods for assessing the magnitudes of the residuals across an entire sample, as well as the likelihood that residuals of that magnitude might have arisen purely by chance are presented below.

☐ **Procedure for estimating magnitude of residuals.** Assume that a sample is chosen consisting of securities with trading events occurring over a certain period of time. For a given event, we define month 0 to be the month in which the event occurred. Similarly, month +1 is the month following the month of the event and so on. Note that months 0, 1, etc., are defined relative to an event and are *not* calendar months. We define U_m, the mean residual for month m over all securities in the sample as

$$U_m = \sum_{i=1}^{N} \left(\frac{\hat{U}_{i,m} \cdot H_i}{N} \right), \qquad (2)$$

where

N = Number of securities in the sample;

$\hat{U}_{i,m}$ = Estimated residual for security i in month m, where m is defined relative to the month of an event;

H_i = 1, if the event of the ith security is a purchase or set of purchases,

= —1, if the event is a sale or set of sales.

Next, define the cumulative average residual, CU_k, as the summation of U_m from month 1 to month k:

$$CU_k = \sum_{m=1}^{K} U_m.^{20} \qquad (2a)$$

The mean residual for month m, U_m, can be interpreted as the average performance across all securities in the mth month following each event. Similarly, the cumulative average residual, CU_k, can be interpreted as the cumulative performance from months 1 to k across all securities in the sample. In this study the first month average residual, the cumulative average residual for the first and the second month, and the cumulative average residual from the first to the eighth month will be examined for each of our samples.

---

in Fama *et al.* [17]. The present study uses the estimates of $\tilde{\gamma}_{0t}$ and $\tilde{\gamma}_{1t}$ derived in Fama and MacBeth [19], which run from January 1935 to June 1968.

[19] *Official Summary of Insider Trading* [46].

[20] For a more detailed explanation of this statistic, see Fama [16] and Fama *et al.* [17].

☐ **Significance tests.** The statistics presented in (2) and (2a) will be used to indicate the magnitude of the residuals in months following insider trading events. We are also interested in testing whether residuals of a particular magnitude could have arisen by chance. Unfortunately, there are certain dependencies among residuals that complicate the specification of a test statistic designed to answer this question.

There is ample evidence that the residual of a security in one calendar month is uncorrelated with the residuals of that security (or any other security) in subsequent calendar months.[21] However, findings presented in King[22] and Meyers[23] strongly suggest that the residuals of different securities in a given calendar month are correlated. As the observations for any given sample are collected over a short period of time, we may be observing the residuals of many different securities in a single calendar month. Since these residuals can be expected to be correlated, a test statistic should not assume independence of these residuals. A statistical test is presented in this section which measures the residual variance of a portfolio of securities directly, thus taking account of the correlation beween residuals of different securities in a given calendar month.

For a sample of companies with insider trading events, we calculate a test statistic indicating whether residuals subsequent to trading events are statistically larger than zero, as follows. First, for each calendar month t covered in our sample, a portfolio of securities is formed. Letting X represent a specified integer,[24] a portfolio corresponding to month t is formed by including the securities of all companies with events between and including month $t - X$ and month t.[25] For a given value of X, there is one portfolio t for each calendar month.[26] If a company has events in c different months during the period from month $t - X$ to month t, it is included c times in portfolio t.

For example, imagine a sample consisting of GM with an event in June 1962, IBM with events in both July 1962 and September 1962, and ITT with an event in October 1962. In Table 1 we list the securities included in each portfolio, given first that $X = 1$ and next that $X = 2$.

The performance of portfolio t during calendar month $t + 1$ is measured as follows:

---

[21] Actually, the most convincing direct evidence of the serial independence of security performance is presented in terms of the returns of securities (see Cootner [11] and Fama [15], in particular). In addition, work by Black and Scholes [5], Ibbotson [26], and unpublished work of the author suggest that residuals are also serially uncorrelated.

[22] King [32].

[23] Meyers [41].

[24] The values of X used in this study are mentioned later.

[25] The procedure for calculating test statistics is most easily explained when t is a calendar month. This is in contrast to the previous section, where time in equations (2) and (2a) is measured relative to an insider trading event.

[26] If there are no events between month $(t - X)$ and month t, portfolio t will contain no securities. As will be seen later, portfolios that contain no securities are ignored in statistical tests.

TABLE 1

SAMPLE OF COMPANIES WITH INSIDER TRADING EVENTS

PORTFOLIO OF	SECURITIES INCLUDED WHEN X = 1	SECURITIES INCLUDED WHEN X = 2
MAY 1962	(NONE)	(NONE)
JUNE	GM	GM
JULY	GM IBM	GM IBM
AUGUST	IBM	GM IBM
SEPTEMBER	IBM	IBM (TWICE)
OCTOBER	IBM ITT	IBM ITT
NOVEMBER	ITT	IBM ITT
DECEMBER	(NONE)	ITT
JANUARY 1963	(NONE)	(NONE)

$$\hat{e}_{t,t+1} = \frac{\sum_{i=1}^{S} (\hat{e}_{i,t+1} \cdot H_i)}{S}, \qquad (3)$$

where

$\hat{e}_{t,\,t+1}$ = Residual of portfolio t in month $t+1$;

$\hat{e}_{i,\,t+1}$ = Residual of the ith security of portfolio t in calendar month $t+1$;

S = Number of securities in portfolio t; and

H_i = 1, if the event of the jth security is a purchase or set of purchases,

= —1, if the event is a sale or set of sales.

An estimate of the variability of the performance of portfolio t, called $\widehat{SD}_t$, is defined as the computed standard deviation of the residual of portfolio t, using data during the period from month (t-49) to month t. This can be written as

$$\widehat{SD}_t = \sqrt{\frac{1}{49} \sum_{j=1}^{50} \left(\hat{e}_{t,t-j+1} - \frac{1}{50} \sum_{i=1}^{50} \hat{e}_{t,t-i+1} \right)^2} \ . \qquad (4)$$

From (3) and (4) we can calculate the standardized residual, $\widehat{se}_{t,\,t+1}$, for portfolio t at time $t+1$ as

$$\widehat{se}_{t,t+1} = \frac{\hat{e}_{t,t+1}}{\widehat{SD}_t}. \qquad (5)$$

This standardized residual, then, is the performance of portfolio t in month $t+1$, deflated by an estimate of the variability of the portfolio's performance in previous months.

As a different portfolio is formed for each calendar month, portfolio t is one of many portfolios. The average standardized residual across all of these portfolios, called $\overline{se}$, is defined as:

$$\overline{se} = \frac{1}{n} \sum_{t=51}^{401} \widehat{se}_{t,t+1} \cdot D_t,^{27} \qquad (6)$$

---

[27] As $\tilde{\gamma}_{0t}$ and $\tilde{\gamma}_{1t}$ are calculated from January, 1935, to June, 1968,

where:

$$D_t = \begin{cases} 1, \text{ when there is at least one security in portfolio } t \\ 0, \text{ when there are no securities in portfolio } t; \end{cases}$$

n = Number of months in which the month's portfolio has at least one security.

It follows that

$$n = \sum_{t=51}^{401} D_t. \tag{7}$$

The standardized residual for portfolio t, $\widehat{se}_{t,\,t+1}$, is an average of many individual residuals observed in month $t + 1$. As residuals in a given calendar month can be expected to be uncorrelated with residuals in other calendar months,[28] the standardized residuals across different months can be treated as independent observations. Under the null hypothesis, $H_0{:}\overline{se} = 0$, the statistic

$$\frac{\overline{se}}{s/\sqrt{n}} \sim t(V), \tag{8}$$

where

$V \approx 49n$ degrees of freedom[29] and

s is the estimate of standard deviation of each standardized portfolio. This value is constrained to be one due to the standardization process.

The above test statistic will be used in this study to determine whether the magnitude of the residuals following a sample of insider trading events could have arisen by chance. Due to the complexity of the procedure, the statistic may appear "unaesthetic" at first glance. However, this complexity is needed. As we know that residuals in a given calendar month are correlated across securities but do not know the exact extent of the correlation, a test statistic is difficult to develop. In our solution, a portfolio is formed in a given calendar month consisting of all the securities whose residuals we observe in the next month. As ex-

---

month 1 is January, 1935 and month 402 is June, 1968 (see note 18). In equation (6), t begins at 51 because 50 months of past data are required to estimate a portfolio's residual variance. Similarly, month 402 is omitted, as one month of future data is required to calculate $\widetilde{se}_{t,\,t+1}$.

[28] *Supra* note 21.

[29] If the residuals of individual securities are normally distributed, the standardized residuals of equation (5) will behave as if drawn from a t-distribution with 49n degrees of freedom. The t-statistic, a weighted sum of these t-distributed variables, is t-distributed with degrees of freedom approximately equal to 49. (See Brownlee [7], pp. 297–304). The degrees of freedom are very large and a normal distribution will suffice. Plots of standardized residuals from samples to be presented later suggest that the individual residuals are indeed t-distributed with a variance of one.

For samples of insider trading to be presented later, the serial correlation of standardized residuals across portfolios was measured. As the null hypothesis that the serial correlation is zero could not be rejected, the individual standardized residuals were treated as independent observations.

plained above, the standardized residual of the portfolio for each calendar month can be treated as an independent observation, implying that the test statistic of (8) is unbiased.

In this study, X assumes the values of zero, one, and seven. Thus, there are three sets of portfolios, each set corresponding to a different value of X. The steps outlined above are repeated for each value of X, yielding three t-statistics. This approach is refered to henceforth as the portfolio method.

3. Performance of insider traders

■ It is meaningful to consider the effect of regulation on the speculation of insiders only if it can first be shown that insiders earn profits above those accruing to investors with no special information. This section presents results suggesting that insiders do earn abnormal profits. First, an initial sample, containing randomly chosen insider trades, is selected as a benchmark. Next, samples consisting of large trades and trades from months of intensive insider activity are studied.

The initial sample is composed of insider trades from the 200 largest securities listed on the Stock Price Tape of the Center for Research in Security Prices (CRSP) of the University of Chicago. For each of these securities, insider trades are observed in five separate months during the interval from 1962 to 1968. The individual months of observation for a particular security are separated in time by 12 to 18 months. The exact number of months between two observations is determined by the selection of a random number from 12 to 18.[30]

In any calendar month an individual trader is classified as a purchaser if the number of days during the month in which he buys stock is greater than the number of days in which he sells stock. Conversely, he is classified as a seller if the number of days in which he buys stock is less than the number of days in which he sells stock. If he purchases stock on just as many days as he sells it, he is not included in the sample. (Exercises of options are excluded from this and all other samples in this study because it is felt that options are exercised due to institutional factors rather than as a result of special information.)

For each company in the initial sample, a month is classified as a month of net purchasers or a month of net sellers depending on whether the number of purchasers is greater or less than the number of sellers. Months during which the number of purchasers equals the number of sellers are excluded. Months of net purchasers and months of net sellers are defined as insider trading *events*. In this and all other samples in the study, events are excluded if companies are not listed on the CRSP tape for fifty months before the event and ten months after the event. This restriction assures sufficient data to form portfolios and to calculate residual variances. There are 952 trades in the initial sample. Many companies have more than one trade in a month, so that the number of months of net purchasers plus the number of months of net sellers is only 362.

---

[30] The exact sequence is contained in Jaffe [27].

Cumulative average residuals are calculated via the method explained in the previous section. In equations (2) and (3), the value of H_t is 1 for a month of net purchasers, while H_t is set to -1 for a month of net sellers. The magnitude of the residuals from this sample, as well as the statistical significance of the residuals, is examined. The average residual of month 1, the cumulative average residual from month 1 to month 2, and the cumulative average residual from month 1 to month 8 are presented in Table 2. These three time intervals are chosen in order to study both the short-term and long-term predictive power of insiders. The cumulative average residual rises to 0.0136 within eight months. This value is of small magnitude, as it would not cover normal transactions costs of two or three percent on a round-trip transaction. Almost one half of that rise, 0.0060, occurs in the first month. This suggests that insiders are better able to predict short-term than longer-term movement in stock price.

In order to test whether the residuals are significantly different from zero, the portfolio method is applied to the sample. Again, the months of net purchasers and the months of net sellers are classified as insider trading events. The results from the portfolio method for the initial sample are also presented in Table 2. For each value of X, the t-value corresponding to the hypothesis that the expected value of the standardized residuals in (8) equals zero, and the one-tailed significance level of that t-value are shown.[31] The t-values appear to be bordering on statistical significance, as two of the three have a significance level under 5 percent.

Though the evidence of the initial sample suggests that insiders possessed and used special information, certain subsamples might yield more convincing results. In particular it was felt that insiders with information might trade large blocks of stock. All transactions from the initial sample whose values are greater than $20,000 are assembled into a subsample. This subsample contains 370 trades, representing 39 percent of the original sample of 952. The previous classification scheme is employed to separate months of net purchasers from months of net sellers. As shown in Table 2, the cumulative average residuals and the t-values are of similar magnitude to the residuals and t-values from the initial sample. These results suggest successful forecasting by insiders, though they do not indicate that large transactions contain more information than other transactions contain.

The work of Glass, Lorie-Neiderhoffer, and Rogoff[32] suggest that months of intensive insider trading precede significant stock price movement. As these studies employed small samples and

---

[31] In this study, X assumes the values of zero, one, and seven. These correspond to periods of one, two, and eight months, within which occurrence of an insider trading event will lead to the inclusion of the corresponding security in a portfolio. Hence, the t-values when X equals zero, one, and seven, respectively, are presented in Table 2 on the same line with the cumulative average residual for one month, two months, and eight months, respectively.

As each t-statistic has a large number of degrees of freedom (see note 29), the one-tailed significance levels are taken from a normal probability table.

[32] *Supra* note 7.

TABLE 2

CUMULATIVE AVERAGE RESIDUALS AND THEIR ASSOCIATED t-VALUES AND SIGNIFICANT LEVELS OVER DIFFERENT PERIODS OF TIME FOR SAMPLES OF INSIDER TRADING

LENGTH OF TIME PERIOD	CUMULATIVE AVERAGE RESIDUAL	CUMULATIVE AVERAGE RESIDUAL PER MONTH	t -VALUE*	ONE=TAILED SIGNIFICANCE LEVEL**
INITIAL SAMPLE - 362 OBSERVATIONS				
ONE MONTH	0 0060	0 0060	1 93	0.026
TWO MONTHS	0.0118	0.0059	2.24	0 012
EIGHT MONTHS	0 0136	0.0017	1 32	0.010
SAMPLE OF LARGE TRANSACTIONS - 204 OBSERVATIONS				
ONE MONTH	0 0062	0.0062	1.99	0 023
TWO MONTHS	0.0134	0.0067	2.09	0.018
EIGHT MONTHS	0.0184	0 0023	1 14	0.126
INTENSIVE TRADING SAMPLE OF THE 1960s WHERE Y EQUALS THREE - 861 OBSERVATIONS				
ONE MONTH	0.0098	0.0098	3 65	0.0001
TWO MONTHS	0 0209	0 0105	4.73	<0 0001
EIGHT MONTHS	0 0507	0.0063	5.23	<0 00001
INTENSIVE TRADING SAMPLE OF THE 1950s WHERE Y EQUALS FOUR - 293 OBSERVATIONS				
ONE MONTH	0 0094	0 0094	3.06	0.001
TWO MONTHS	0 0174	0 0087	3.16	0 0008
EIGHT MONTHS	0 0514	0 0064	4 69	<0.00001
INTENSIVE TRADING SAMPLE OF THE 1950s WHERE Y EQUALS FIVE-157 OBSERVATIONS				
ONE MONTH	0.0112	0.0112	2.97	0.001
TWO MONTHS	0 0174	0 0087	2.76	0.003
EIGHT MONTHS	0 0448	0 0056	3 18	0 007
INTENSIVE TRADING SAMPLE OF THE 1950s WHERE Y EQUALS SIX - 80 OBSERVATIONS				
ONE MONTH	0.0125	0 0125	1 60	0 055
TWO MONTHS	0 0078	0 0010	0.77	0.227
EIGHT MONTHS	0.0360	0 0045	1 05	0 147
INTENSIVE TRADING SAMPLE FROM APRIL TO OCTOBER 1961-131 OBSERVATIONS				
ONE MONTH	0.0164	0 0164	2.31	0.0104
TWO MONTHS	0 0236	0 0116	2.00	0 0224
EIGHT MONTHS	0 0848	0 0106	2 88	0 0019
INTENSIVE TRADING SAMPLE FROM DECEMBER 1961 TO NOVEMBER 1962 - 203 OBSERVATIONS				
ONE MONTH	0.0093	0 1096	1 44	0.0749
TWO MONTHS	0 0196	0 0098	2.13	0.0166
EIGHT MONTHS	0 0512	0 0064	2 18	0.0146
INTENSIVE TRADING SAMPLE FROM JANUARY 1964 TO MARCH 1965-278 OBSERVATIONS				
ONE MONTH	0.0053	0 0053	1.31	0.0951
TWO MONTHS	0.0166	0.0083	2.42	0.0078
EIGHT MONTHS	0.0376	0 0047	2 47	0 0068
INTENSIVE TRADING SAMPLE FROM MAY 1965 TO DECEMBER 1965-136 OBSERVATIONS				
ONE MONTH	0.0160	0 0160	2.72	0.0033
TWO MONTHS	0.0340	0.0170	3.37	0 0004
EIGHT MONTHS	0 0416	0 0052	1 97	0.0244
INTENSIVE TRADING SAMPLE FROM SEPTEMBER 1966 TO MARCH 1967-112 OBSERVATIONS				
ONE MONTH	0 0076	0 0076	0 88	0 1894
TWO MONTHS	0.0150	0.0075	0 94	0 1736
EIGHT MONTHS	0.0840	0 0105	1.97	0 0244

*FOR EACH SAMPLE, THE THREE t-VALUES ARE FOR X = 0, 1, 7 RESPECTIVELY (SEE NOTE 31).

**BASED ON NORMAL DISTRIBUTION (SEE NOTE 29) ignored the relative risk of different securities and general market conditions, this paper also examines intensive trading months. Any company with at least Y more purchasers than sellers in a month is classified as an intensive buying company for that month. Conversely, any company with at least Y more sellers than purchasers in a month is classified as an intensive selling company for the month. In the following sample, Y will assume the value of three.

The present sample includes all intensive trading companies listed on the CRSP tape during the months from April to October 1961; from December 1961 to November 1962; from January 1964 to March 1965; from May to December 1965; and from September 1966 to March 1967.[33] Using the methods explained in Section 2, the cumulative average residuals and t-values are calculated. In equations (2) and (3), the value of H_i is 1 for an intensive buying month, while H_i is set at -1 for an intensive selling month. The cumulative average residuals and t-values are presented in Table 2. The residuals rise approximately 5 percent in eight months, with 3 percent of the rise occurring in the last six months. These figures are quite large, as they are much greater than transaction costs. The t-values are clearly significant as they range from 3.65 to 5.23. The results from this sample support the previous research cited above.

The above results were checked by selecting a sample of intensive trading companies from a different time period. For this sample Y assumes the values of four, five, and six. All intensive trading months occurring during the period from January 1953 to December 1955 and January 1958 to December 1959 for the companies listed on the CRSP tape are included in the sample. The results from these samples are also presented in Table 2.

For the sample when Y equals four, the cumulative average residuals rise over 5 percent in eight months, suggesting that insiders possess special information. The t-values are large, as each of the three is greater than 3.0. The residuals when Y equals four are lower than the t-values in the former sample due to a smaller sample size. When Y equals six, the residuals are smaller and the t-values are not significant. As this sample contains only eighty observations, the results are not so decisive as those of the total sample.

The findings from these intensive trading samples also suggest that insiders are earning abnormal profits. However, as the cumulative average residuals do not increase as Y increases, the results do not suggest that the profit of insiders is an increasing function of the intensity of the trading.

The study has examined three different samples: the initial sample and two independent samples of intensive trading companies. Each of these samples exhibits evidence of special information among insiders. The sample of intensive trading companies yields particularly strong evidence of special information as residuals from some of the subsamples rose above 5 percent in eight months, while some t-values were above five. Because evidence has been provided that insiders can earn significant profits, there is now reason to examine the effect of security regulation on the profits of insiders.

---

[33] The sample contains insider trades immediately before and after three regulatory changes. Subsamples will be used in Section 4.

4. Effectiveness of insider trading regulation

■ The effect of regulation on the securities market is a subject of much current interest. For the most part, it has simply been assumed that regulation accomplishes its objectives. Only rarely is it possible to directly measure the effects of a particular type of governmental control. The regulation of insiders is one of these exceptions. Since we have seen in the previous section that insiders do possess special information, we can now compare the extent of their use of that information in various periods of time with the degree of regulation in effect in these periods.

As previously mentioned, three of the most significant changes in security regulation have resulted from the Cady, Roberts decision, the Texas Gulf Sulphur indictment, and the Texas Gulf Sulphur decision. Though the Securities and Exchange Act of 1934 empowered the SEC with authority over insiders, this authority was not enforced until the Cady, Roberts decision in November, 1961. The decision penalized a brokerage firm for trading with secret information on behalf of its clients and thus established the precedent that corporate officials trading on inside information were liable for civil prosecution. In April 1965 the SEC indicted officials of Texas Gulf Sulphur for trading on, as well as suppressing, information related to the company's vast mineral strike. Since the Cady, Roberts decision was merely an administrative ruling, the Texas Gulf Sulphur civil indictment can be viewed as an extension of regulatory law. To the surprise of many lawyers and corporate officials, the decision in August 1966 upheld most of the counts of the April 1965 indictment. This ruling showed insiders that the courts, as well as the SEC, would punish speculation on special information. In recent years, the courts have handled other cases involving misuse of inside information, and have heard appeals on the Texas Gulf Sulphur case.[34] However, since the Cady, Roberts decision and the Texas Gulf Sulphur indictment and decision appear to be the most important, the effects of only these three actions are examined.[35] In this section, we compare the properties of insider trading immediately before and after each of these regulatory changes. Since a regulatory decision could conceivably affect both the profitability and the volume of insider trades, the existence of these possible effects is investigated below.

☐ **Effect on profitability.** Prior research[36] and the results of Section 3 suggest that an examination of intensive trading companies yields the most convincing evidence of successful forecasting by insiders. Below we compare the performance of intensive trading companies prior to each decision with the performance of intensive trading companies subsequent to it. For this section of the study, any company with at least three more purchasers than sellers in a month is classified as an intensive buying company for that month, i.e., we set $Y = 3$. Conversely, any company with at least three more sellers than purchasers in a month is classified as an intensive selling company.

---

[34] For example SEC v. Texas Gulf Sulphur Co., 401 F. 2d. 833 (2d Cir. 1968) and SEC v. Texas Gulf Sulphur, 312 F. Supp. 77 (S.D.N.Y., 1970). For a sample of the class action suits filed against Texas Gulf Sulphur, see Reynolds v. Texas Gulf Sulphur, 309 F. Supp. 548 (D. C. Utah, 1970).

[35] The three cases are discussed in greater detail in the appendices.

[36] *Supra* note 7.

As the Cady, Roberts decision occurred in November 1961, we select a sample including every company listed on the CRSP tape with an intensive trading month during the period from April 1961 to October 1961, and another sample including every company listed on the CRSP tape with an intensive trading month during the period from December 1961 to November 1962. We use the methods explained in Section 2 and calculate the cumulative average residuals and t-values for the two samples. In equations (2) and (3) the value of H_i is 1 for an intensive buying month, while H_i is set at -1 for an intensive selling month. The cumulative average residuals and their associated t-values for the two samples are presented in Table 2. In the first sample, the cumulative average residuals are large, as they rise approximately 0.08 in eight months following trading events. Insiders appear to be forecasting returns many months ahead, as the value of the cumulative average residuals increases when we move from one to eight months. However, as the cumulative average residuals *per month* decline, one suspects that insiders are better able to predict stock price movement in the immediate future than in the more distant future. All of the t-values for the pre-Cady, Roberts sample are larger than 2.00, indicating with high probability that insiders earned excess returns during this period. The cumulative average residuals are large in the second sample and the same general pattern of increasing cumulative average residuals but decreasing cumulative average residuals per month is observed here. We also find high t-values in this sample. However, the cumulative average residuals are lower than in the pre-Cady, Roberts sample, possibly indicating an effect due to the decision.

To examine the hypothesis that the Cady, Roberts decision had no effect on the profitability of insider trading, a difference of the means test is constructed. For each value of X, the average standardized residual across portfolios from the pre-Cady, Roberts data is tested to be significantly different from the average standardized residual[37] from the post-Cady, Roberts sample.

The t-statistic for this test can be written as follows:

$$t_x = \frac{\overline{sr}_{2x} - \overline{sr}_{1x}{}^{[38]}}{s\sqrt{1/n_{1x} + 1/n_{2x}}}, \qquad (9)$$

where

---

[37] The definition for average standardized residual is presented in Section 2.

[38] See Brownlee [7], p. 298.

$\overline{sr}_{1x}$ = Average standardized residual for pre-Cady, Roberts data for a particular value of X;

$\overline{sr}_{2x}$ = Average standardized residual for post-Cady, Roberts data for a particular value of X;

n_{1x} = Number of pre-Cady, Roberts portfolios for a particular value of X;

n_{2x} = Number of post-Cady, Roberts portfolios for a particular value of X;

s = Estimate of standard deviation. The standardization process sets this value at one.

The number of degrees of freedom is very large so the normal approximation suffices.[39]

The t-values associated with the three difference of the means tests are presented in Table 3. All three t-values are negative, indicating greater insider profits prior to the decision. However, the t-values are statistically small with the most negative of the three

TABLE 3

TESTS OF WHETHER RESIDUALS FOLLOWING TRADES OF INSIDERS PRIOR TO A REGULATORY CHANGE DIFFER SIGNIFICANTLY FROM TRADES SUBSEQUENT TO THE CHANGE

X	t-VALUE	ONE – TAILED SIGNIFICANCE LEVEL+
THE CADY, ROBERTS DECISION*		
0	−1.04#	0.1492
1	−0.60	0.2743
7	−0.21	0.4168
THE TEXAS GULF SULPHUR INDICTMENT**		
0	1.35	0.9115
1	1.44	0.9251
7	−0.07	0.4761
THE TEXAS GULF SULPHUR DECISION***		
0	−1.26	0.1038
1	−1.76	0.0392
7	+0.05	0.5200
ALL THREE DECISIONS****		
0	−0.86	0.1949
1	−0.74	0.2296
7	−0.69	0.2451

+BASED ON THE NORMAL DISTRIBUTION (SEE FOOTNOTE 29)

#A NEGATIVE t-VALUE INDICATES THAT THE AVERAGE STANDARDIZED RESIDUAL FOR THE SAMPLE PRIOR TO A REGULATORY CHANGE IS LARGER THAN THE AVERAGE STANDARDIZED RESIDUAL SUBSEQUENT TO THE CHANGE

*THE SAMPLE FROM APRIL 1961 TO OCTOBER 1961 IS COMPARED WITH THE SAMPLE FROM DECEMBER 1961 TO NOVEMBER 1962.

**THE SAMPLE FROM JANUARY 1964 TO MARCH 1965 IS COMPARED WITH THE SAMPLE FROM MAY 1965 TO DECEMBER 1965

***THE SAMPLE FROM MAY 1965 TO DECEMBER 1965 IS COMPARED WITH THE SAMPLE FROM SEPTEMBER 1966 TO MARCH 1967.

****THE SAMPLE FROM APRIL 1961 TO OCTOBER 1961 IS COMPARED WITH THE SAMPLE FROM SEPTEMBER 1966 TO MARCH 1967 being —1.04. As values this low imply that the differences might well be due to chance, the hypothesis that the Cady, Roberts decision did not affect the profitability of insider trading cannot be rejected.

---

[39] *Supra* note 29.

In order to measure the effects of the Texas Gulf Sulphur indictment on the profitability of insider trading, two samples are used. The first sample includes all intensive trading months during the period from January 1964 to March 1965 for any company listed on the CRSP tape, while the second sample includes all intensive trading months during the period from May to December, 1965.[40] For each of the two samples, the results from residual analysis and the portfolio method are presented in Table 2. The cumulative average residuals are large in these samples, though smaller than those from the Cady, Roberts samples. All six t-values are large and positive, suggesting that insiders were earning profits both before and after the Texas Gulf Sulphur indictment. It is interesting that insiders actually perform slightly better *after* the indictment, as all three cumulative average residuals are higher in the post-TGS sample.

To examine the null hypothesis that insiders were not influenced by the Texas Gulf Sulphur indictment, the average standardized residuals from the pre-TGS sample are tested against the average standardized residuals from the post-TGS sample. The results of the difference of the means tests between the two samples are presented in Table 3. The findings here are mixed. The first two of the three t-values are positive and of moderate size. This indicates that insiders earned higher profits after the decision. In contrast, the third t-value is slightly negative, suggesting that the long-run performance of the two samples is nearly equal.[41] More important, as none of the t-values is statistically large, there is not sufficient evidence to reject the hypothesis that the Texas Gulf Sulphur indictment did not affect the profitability of insider trades.

Though the Texas Gulf Sulphur indictment did not appear to have a deterrent effect on the profitability of insider trading, the decision on August 19, 1966, may have had such an effect. To test this, a sample containing all the intensive trading events of companies listed on the CRSP tape from September 1966 to March 1967 is collected. Statistics from residual analysis are presented in Table 2. The cumulative average residuals from this sample are as large, if not larger, than those from prior samples. The eight-month value of 0.0840 is particularly high, suggesting that the three regulatory decisions may have had only a slight, if any, effect on the performance of insiders. This sample is compared to the one running from May to December, 1965. (Statistics on this latter sample are also presented in Table 2.) The one- and two-month cumulative average residuals are larger in the 1965 sample than in the 1966 to 1967 sample, while the eight-month cumulative average residual is smaller.

---

[40] Many companies with intensive trading in April, 1965, the month of the Texas Gulf Sulphur indictment, exhibit trading both before and after April 19, the date of the indictment. It is unclear whether these companies should have been included in the pre-Texas Gulf Sulphur (pre-TGS) or the post-Texas Gulf Sulphur (post-TGS) sample. Thus, April 1965 is excluded from the two samples.

[41] At first glance, it may appear contradictory that the eight-month cumulative average residual of the pre-TGS sample is smaller than the eight-month cumulative average residual of the post-TGS sample, while the t-value of -0.067 is negative. However, due to the standardization process, the standardized residual when $X = 7$ of the pre-TGS sample is actually slightly larger than the standardized residual of the post-TGS sample, implying a negative t-value.

The results of the difference of the means tests between the average standardized residuals of the two samples are presented in Table 3. Again, the findings are mixed. The first two *t*-values are negative, indicating a decrease in insider profitability after the decision, and when *X* equals one, the *t*-value is —1.76, a figure bordering on statistical significance. This evidence suggests that the Texas Gulf Sulphur decision may have had an effect on insider trading. However, the positive third *t*-value implies that the long-run trading gains from the second sample are actually greater.

The cumulative effect of the three cases is measured by comparing the pre-Cady, Roberts decision data with the post-Texas Gulf Sulphur decision data. From Table 2, we find that the one- and two-month cumulative average residuals are higher in the pre-Cady, Roberts data, while the eight-month cumulative average residuals are almost identical in the two samples. The results of difference of the means tests between the average standardized residuals of the pre-Cady, Roberts sample and the average standardized residuals of the post-Texas Gulf Sulphur sample are presented in Table 3. The *t*-values are negative, indicating better performance in the pre-Cady, Roberts sample. However, since the *t*-values are small (all are above —1.0), the evidence is not sufficient to reject the hypothesis that there was no cumulative effect of the three events.

In this section, we have examined the individual effects of three regulatory events on the profitability of insider trades. On the one hand, there was a drop in the average profitability of insider trades following both the Cady, Roberts decision and the Texas Gulf Sulphur decision. On the other hand, there appeared to be an increase in profitability following the Texas Gulf Sulphur indictment. However, for each of the three regulatory events, the average profitability of insider trades before an event was not significantly different from the average profitability of insider trades after that event. In addition, the performance of insiders in a period prior to the three events was not significantly different from the performance of insiders in a period subsequent to all three events. Hence, the results of this section do not suggest that regulation changes had an effect on the profitability of insider trading.

☐ **Effect on insider trading volume.** As the total profit to insiders per period might be related to the frequency of trades, the trading activity in the months surrounding our three case law changes is examined. A sample is selected containing the total number of insider traders in months before and after each regulation change. Options are excluded in the formation of the sample. To detect an effect from a decision, a time series model of the number of traders per month could be specified so that the level of the residuals from the model prior to the decision could be tested against the level of the residuals subsequent to the decision.

In examining an effect due to the Cady, Roberts decision, the number of traders is calculated for each month from May 1961 to May 1962. There is an average of 821 insider traders per month in the six months from May to October, 1961. Trading frequency actually increased in the six months from December 1961 to May 1962, as there is an average of 846 traders per month here. The variance of the monthly series running from May 1961, to May 1962, is presented in Table 4. The difference in the number of traders is both small compared to the variance of the series and positive so that no explicit time series model is needed.

TABLE 4

STATISTICS ON THE NUMBER OF INSIDER TRADERS AND STANDARDIZED TRADERS IN SPECIFIC CALENDAR MONTHS

PERIOD	AVERAGE NUMBER PER MONTH	ESTIMATED STANDARD DEVIATION	MONTHS USED TO ESTIMATE STANDARD DEVIATION
INSIDER TRADERS			
MAY-OCT 1961	821	102.4	MAY 1961-MAY 1962
DEC 1961-MAY 1962	846	102.4	MAY 1961-MAY 1962
OCT 1964-MARCH 1965	1020	174.6	OCT 1964-OCT 1965
MAY 1965-OCT 1965	1228	174.6	OCT 1964-OCT 1965
FEB 1966-JULY 1966	1040	206.1	FEB 1966-FEB 1967
SEP 1966-FEB 1967	1136	206.1	FEB 1966-FEB 1967
STANDARDIZED TRADERS			
MAY-OCT 1961	846	125.1	MAY 1961-MAY 1962
DEC 1961-MAY 1962	844	125.1	MAY 1961-MAY 1962
OCT 1964-MARCH 1965	1077	133.5	OCT 1964-OCT 1965
MAY 1965-OCT 1965	1153	133.5	OCT 1964-OCT 1965
FEB 1966-JULY 1966	1039	210.2	FEB 1966-FEB 1967
SEP 1966-FEB 1967	1206	210.2	FEB 1966-FEB 1967

To check for immediate effects, a time series of the first differences of the number of traders in successive calendar months is formed. Since one might expect a once-and-for-all drop in the number of traders following a decision, the first differences in the month of the decision and the following month are examined.[42] Should the first differences in these two months fall far below those of surrounding months, a time series model from all the first differences would be specified. In this way, the first differences of these two months could be tested to determine whether they are significantly different from the first differences of other months. In Table 5, the mean and standard deviation from this series are

---

[42] For example, as the Cady, Roberts decision occurred November 8, 1961, a once-and-for-all drop in trading would result in reduced insider activity from November 9th on. Hence, we would expect a low volume of trading in November due to the effect of the decision on the last 22 days of the month. However, we would expect an even lower volume of trading in December as all of its 31 days would be affected. Therefore, a once-and-for-all drop in trading due to the decision would yield negative first differences in both November and December.

TABLE 5

STATISTICS ON THE FIRST DIFFERENCES OF THE NUMBER OF INSIDER TRADERS AND STANDARDIZED TRADERS IN SPECIFIC CALENDAR MONTHS

MONTH OF FIRST DIFFERENCE	FIRST DIFFERENCE	ESTIMATED STANDARD DEVIATION FROM SURROUNDING MONTHS	ESTIMATED MEAN FROM SURROUNDING MONTHS	SURROUNDING MONTHS USED TO ESTIMATE MEAN AND STANDARD DEVIATION OF THE SERIES OF FIRST DIFFERENCES
INSIDER TRADERS				
NOVEMBER 1961	+58	160 59	7.41	MAY 1961-MAY 1962
DECEMBER 1961	+90	160.59	7 41	MAY 1961-MAY 1962
APRIL 1965	−150	86.35	25 89	OCT 1964-OCT 1965
MAY 1965	+7	86.35	25.89	OCT 1964-OCT 1965
AUGUST 1966	+451	343 79	9 67	FEB 1966-FEB 1967
SEPTEMBER 1966	−385	343 79	9.67	FEB 1966-FEB 1967
STANDARDIZED TRADERS				
NOVEMBER 1961	−87	186.63	9.77	MAY 1961-MAY 1962
DECEMBER 1961	+114	186 63	9.77	MAY 1961-MAY 1962
APRIL 1965	−93	184 48	2.11	OCT 1964-OCT 1965
MAY 1965	+84	184.48	2.11	OCT 1964-OCT 1965
AUGUST 1966	+13	315 81	4.42	FEB 1966-FEB 1967
SEPTEMBER 1966	−3	315.81	4 42	FEB 1966-FEB 1967

compared with the first differences in November and December, 1961. As the first differences of these two months are small relative to the standard deviation, no time series is specified.

In the above procedure we implicitly assumed that the trading frequency of insiders could only be affected by a regulatory decision. However, general market conditions may also influence trading activity. The total monthly volume on the New York Stock Exchange (NYSE) is used as a proxy for general market conditions.[43] In order to neutralize the effect of market volume, we divide the number of traders in a month by a factor proportional to the total monthly volume in shares on the NYSE. To calculate this number of "standardized" traders, we define X_t, the ratio of the number of insider traders to monthly market volume in month t, as

$$X_t = \frac{Y_t}{W_t}, \tag{10}$$

where

Y_t = Number of insider traders in month t; and

W_t = Total monthly volume in shares traded on NYSE in month t.

Next we can represent S_t, the number of standardized traders in month t, as

$$S_t = \frac{\sum_{t=1}^{N} Y_t}{\sum_{t=1}^{N} X_t} \cdot X_t, \tag{11}$$

---

[43] These data were collected from the *Bank and Quotation Record* [3].

where N is the number of months examined. Thus we have

$$\sum_{t=1}^{N} S_t = \sum_{t=1}^{N} Y_t.$$

As presented in Table 5, the first differences in November and December, 1961, of the standardized series are —87 and +114, respectively. These again are small compared to the estimated standard deviation and are of opposite signs. An examination of longer term effects showed that in the first six months from May to October, 1961, there are 846 standardized traders per month, while in the six months directly following the month of the Cady, Roberts decision there are 844 standardized traders per month. As can be seen from Table 4, this difference is quite small in comparison with the series' standard deviation.

The above results suggest that the Cady, Roberts decision had no important effect on the amount of trading among insiders. Any changes in the levels of trading or the first differences of the series are small compared to the standard deviations of these series. Thus, the evidence is not sufficient to reject the null hypothesis that the Cady, Roberts decision did not affect either the profitability or volume of insider trading.

The effect of the Texas Gulf Sulphur indictment on the volume of insider trading is examined. We compare trading frequency in the months immediately before April 1965, the month of the indictment, with the frequency in the months immediately following April. From October 1964 to March 1965 there is an average of 1020 traders per month, while from May to October, 1965, there is an average of 1228 traders per month. Thus trading again increased after the regulatory event. This disparity is reduced through standardization, as the two standardized figures are 1077 and 1153, respectively. As can be seen from Table 4, the difference in the standardized series is small in comparison to the standard deviation calculated from the monthly series of standardized traders from October 1964 to October 1965. Both the series of the unstandardized traders and the series of the standardized traders yield a positive and negative first difference for the months of April and May, 1965. These first differences are small in comparison to the standard deviation of the series of first differences calculated from October 1964 to October 1965.

The data presented for total traders indicate a slight increase in insider activity after the Texas Gulf Sulphur indictment. As all differences are small compared to sample variation, the evidence from monthly data is not sufficient to reject the null hypothesis that the Texas Gulf Sulphur indictment had no effect on the volume of insider trading.

An examination of daily data may yield noticeable changes in trading.[44] The average number of insider trades per day from

---

[44] In 1963, the *Official Summary of Insider Trades* [46] began listing the month, day and year of any insider trade. However, prior to 1963, only the month and year of each transaction were given. Hence, we can investigate any daily changes in trading activity due to the Texas Gulf Sulphur indictment and the Texas Gulf Sulphur decision, though we were only able to examine the monthly effects of the Cady, Roberts decision.

April 1, 1965, to the indictment on April 19, the number of trades on the day of the indictment, and the average number of trades per day from the indictment to the end of the month are presented in Table 6. The number of trades on the day of the indictment, 75, is larger than the average number of trades per day in the preceding days of the month, 73.3, and the average number of trades per day in the rest of the month, 78.3, is even higher. Hence the findings from daily data do not suggest a deterrent effect due to the indictment.

TABLE 6

NUMBER OF INSIDER TRANSACTIONS DURING MONTHS OF REGULATION CHANGES

DATES	NUMBER OF TRANSACTIONS	NUMBER OF TRADING DAYS	AVERAGE NUMBER OF TRADES PER DAY
APRIL 1— APRIL 18, 1965	880	12	73.3
APRIL 19, 1965	75	1	75.0
APRIL 20— APRIL 30, 1965	626	8	78.3
AUGUST 1— AUGUST 18, 1966	1177	14	84.1
AUGUST 19, 1966	94	1	94.0
AUGUST 20— AUGUST 30, 1966	716	8	89.5

Next, the effect of the Texas Gulf Sulphur decision on the volume of trading is examined. As shown in Table 4, in the first six months prior to the decision, the average number of traders is 1040 per month, while in the first six months subsequent to the decision, the average number of traders per month is 1136. Again volume has increased after a regulatory event. The disparity is even greater for the standardized series, suggesting that the Texas Gulf Sulphur decision did not deter insiders from trading. The first differences for August and September are large for both the standardized and unstandardized time series, but they are of opposite signs and of the same magnitude in both series. Thus, the tables do not suggest evidence sufficient to reject the hypothesis that the decision had no effect on the volume of insider trading.

Daily data surrounding the decision are also presented. Table 6 presents the trading per day for three time periods during August 1966. Trading per day is actually higher after the decision and is even larger on the day of the decision. Thus, daily data do not suggest an effect due to the decision.

In this section, the effect of the three decisions on the volume of trading was investigated. Initially, monthly data were used. Time series of the number of insider traders per month were formed and the data were standardized by total monthly market volume on the NYSE. First months, as well as lagged effects of each decision, were examined. Both the standardized and unstandardized time series did not indicate an effect on volume from any of the three decisions. Daily data were also employed for the last two regulatory events. The results did not indicate any effect due to either event, as the insider trading volume actually increased in the days following the indictment and the decision. Thus, the data do not suggest that the regulation changes in the 1960s influenced the volume of insider trading.

5. Conclusion

■ In order to assess the effectiveness of changes in the regulation of insiders, this study examined certain properties of insider trading immediately before and after three important events in case law. There appeared to be few changes in the characteristics of trading attributable to the three cases. Only the Texas Gulf Sulphur decision seems to have had even a slight effect on the profitability of insider trading, and there is no evidence of any cumulative effect of the three events on profitability. Furthermore, the data do not suggest that the regulatory changes affected the volume of insider trading. The volume actually increased slightly after each of the three events, though none of these changes in volume could be called statistically large.

The nature of the regulations and their enforcement may explain why their effect is so small. Though new developments have increased the restrictiveness of the old laws, these laws still prohibit only the most flagrant examples of speculation on inside information. In order to prosecute an insider, the SEC must prove not only that the individual made abnormal profits, but that he was also trading on "material" information not known to the general public.[45] We previously showed in our intensive trading samples that, on average, insiders reaped excess returns of approximately 4 percent per trade. Though insiders earning these small returns may actually be receiving material information, movements in stock prices of this order are likely to be due to events undetectable by a regulatory agency.

Furthermore, the SEC appears to use superficial investigatory techniques. For example, Manne states, "Even the SEC, which has quite efficient discovery devices at its command, still relies largely on newspaper accounts of specific events to know where to begin a search."[46] Since the detection of violations is difficult, insiders may have assumed that special information must be very conspicuous to warrant civil action. And, the record does suggest that the government indicts only the most obvious violators of the law. For example, the Cady, Roberts case dealt with an individual who purchased stock minutes after learning of an announcement to reduce dividends, and the Texas Gulf Sulphur case investigated trading in the months prior to the tripling of stock price. In addition, punishment does not seem a sufficient deterrent to insiders.

---

[45] The prevailing legal interpretation of materiality stems from the decision of Kardon v. National Gypsum Co. 73 F. Supp. 798,800 (E.D. Pa., 1947) which states that material facts "would materially affect the judgment of the other party to the transaction." This rather vague statement has been interpreted by Jennings and Marsh, Jr. [31], p. 146, who write, "More recently, the courts have also stated the standard in terms of whether in reasonable and objective contemplation the information might affect the value of the corporations' securities."

[46] See Manne [37], p. 163.

The typical outcome of successful prosecution is merely recovery of profits by the corporation, though recent developments in the Texas Gulf Sulphur case could alter this. The maximum fine under the Securities and Exchange Act is \$10,000,[47] a small cost in comparison with the size of many transactions.

The results of this paper do not suggest that the recent regulation of insiders is effective, casting doubt on the value of this regulation to society. However, before policy decisions can be made, much new research is needed in this area. This paper has attempted only to measure the power of current insider regulation. Future work might investigate the factors impeding effective regulation and suggest new techniques to increase the influence of our laws. In particular, methods to restrict the trading of tippees[48] would be valuable. Furthermore, an examination of the costs to taxpayers of security regulation would be both important and interesting.

More generally, insider trading regulation should be related to discussions of social welfare. Though the important works of Demsetz, Fama-Laffer, Friend, Hirshleifer, Manne, and Marshall[49] have gone far in this regard, certain questions are not yet satisfactorily answered. For example, are we interested in the fairest solution or the most efficient one? How much money is it worth to society to prevent the redistribution of a dollar from naive investors to insiders? When these and other questions have been answered, empirical work can be better put into perspective.

Appendix 1

Laws and regulations governing insider trading

Prior to the Securities and Exchange Act of 1933–1934, regulations governing the trading of securities were based on common law. However, the common law regulations have been of little importance in transactions over exchanges. "For example, no case has ever found an insider liable for a transaction that took place over an organized stock exchange. This seems to be an implicit recognition by common-law courts that specific regulation is needed only in those instances when an active market for corporate securities does not guarantee to shareholders sufficient protection against fraud. A leading case recognized this fact implicitly when it stated that mere silence does not usually amount to a breach of duty."[50]

The Securities and Exchange Act was designed to regulate the trading of securities on organized exchanges. Section 16(a) of the Act requires corporate officers, directors, and owners of more than ten percent of any type of security to report transactions in their own stock to the Securities and Exchange Commission (SEC).

---

[47] See [9], pp. 372–376.

[48] A tippee is usually considered to be an individual who receives information from an insider though, according to SEC regulations, he is not an insider himself. See [8] for an in-depth explanation.

[49] See Demsetz [12, 13], Fama-Laffer [18], Friend [23], Hirshleifer [25], Manne [38], and Marshall [40].

[50] See Manne [38], p. 22.

These reports must be received by the SEC ten days after the end of the month of trading and are usually published in the Official Summary approximately five weeks after the end of the month.

Section 16(b) requires that all "short-term" profits by insiders must be returned to the corporation. Short-term profits result from a purchase and subsequent sale within six months. The ambiguity in determining exactly what transactions can be prosecuted under Section 16(b) has resulted in complex litigation, but this need not concern us. The relevant fact is that this law is strictly interpreted. "An insider who holds for six months and a day before selling does not become liable under Section 16(b), no matter how much proof is adduced of unfair resort to nonpublic information."[51]

Section 17(a) and Section 10(b) were enacted to prosecute the manipulations that elude Section 16. Section 17(a) states that fraud or "any omission to state a material fact necessary in order to make the statements made, in the light of circumstances under which they are made, not misleading"[52] shall be outlawed in the sales of securities. However, the wording in this section implies no judgment on purchases and "the Commission has never sought to apply Section 17(a) except against fraudulent sellers."[53]

Section 10(b) of the Securities and Exchange Act declares unlawful "any manipulation, deceptive device or contrivance in contravention of such rules and regulation as the Commission may prescribe."[54] This section applies to both purchases and sales. Section 10(b) was not employed for many years as it is "not operative in the absence of specific rules."[55]

Since Section 17(a) only permits prosecution of sales, the Securities and Exchange Commission encouraged Congress to amend the section to include purchases. When Congress declined, the SEC created rule X-10(b)-5 in 1942 as allowed under Section 10(b). This rule, now called rule 10(b)-5, uses similar language to Section 17(a) but can prosecute purchases as well as sales. "This was the first move by the SEC to reach the purchaser of securities in the type of situation involved."[56]

Though the laws and rules cited above indicate that lack of disclosure in a transaction is a violation, many legal experts contend that the law on nondisclosure was not enforced until the historic Cady, Roberts[57] decision. "Since the adoption of Rule 10(b)-5 in 1942, a question has always existed as to whether the omission to make any disclosure, as distinguished from the use of half-truths, was a violation of the rule. It has been suggested that the Cady, Roberts decision was the first to hold that total nondisclosure was a violation."[58] Hence, 1934 to 1961 may be viewed as a period in which regulations restricting the use of special information by insiders were not enforced.

---

[51] See Loss [35], p. 564.

[52] *Ibid.*, p. 808.

[53] *Ibid.*, p. 808.

[54] *Ibid.*, p. 808.

[55] *Ibid.*, p. 809.

[56] Comment [10], p. 770.

[57] In the matter of Cady, Roberts and Co., SEC Release No. 6668, CCH Federal Securities Law Reporter par 76,803 (1961).

[58] Jennings and Marsh [31], p. 142.

As can be seen, the wording in the sections and rules cited above are vague. Thus, case law must determine exactly what interpretations the legal profession and insiders attach to the sections. Discussion of specific cases follows.

Appendix 2

In the matter of Cady, Roberts and Company

■ On November 8, 1961, the SEC issued a release, In the Matter of Cady, Roberts and Co., which extended existing insider trading regulation. The release charged that Cady, Roberts and Co. and Robert M. Gintel, a broker of Cady, Roberts, "willfully violated antifraud provisions of Section 10(b), Rule 10(b)-5 and Section 17(a)."[59]

In early November, 1959, Curtiss-Wright Corp. notified Robert M. Gintel of a public presentation on November 23 of their new engine. Since Mr. Gintel's invitation was one of only 2,000 received at an early date, the SEC considered it to be new information to the market. That the stock rose from 32 to 35¼ on November 24 was cited as evidence of this new information. From November 6 to the day of the presentation, Mr. Gintel purchased 11,000 shares of Curtiss-Wright stock for his clients. From November 24 through 11 AM on November 25, Mr. Gintel sold 6,500 shares of stock.

The directors of Curtiss-Wright met on November 25 to approve a reduced dividend for the following quarter. At 11 AM the directors decided to release this information to the New York Stock Exchange via telegram. Western Union received the message at 11:12 AM but, because of an unexplained delay, was unable to forward the message to the NYSE until 12:29 PM. A representative of Cady, Roberts and Co. attended the director's meeting and promptly relayed the information concerning the dividend cut to Mr. Gintel. By 11:18 AM, Mr. Gintel had sold another 7,000 shares (including short sales) for his customers. The stock fell 4½ points between 11:18 AM and the close of the day.

Both Mr. Gintel and Cady, Roberts and Co. were found to have violated Section 10(b), Section 17(a), and Rule 10(b)-5. Gintel was fined $3,000 and suspended from the New York Stock Exchange for twenty days.

The Cady, Roberts and Co. decision was significant for two reasons. First, it extended liability to persons other than corporate officers, directors, and beneficial owners. The release stated that the obligation under the antifraud provisions "rests on two principal elements: first, the existence of a relationship giving access, directly or indirectly, to information intended to be available only for a corporate purpose and not for the personal benefit of anyone, and second, the inherent unfairness involved where a party takes advantage of such information knowing it is unavailable to those with whom he is dealing."[60]

Second, it clarified the antifraud provisions of the Securities and Exchange Act. "It laid to final rest the belief that Rule 10b-5 was still circumscribed by the older special facts approach."[61] Though in certain cases, insiders could be prosecuted for trading on special information under the special facts rule in common law, the rule had never been applied to transactions over an organized exchange.[62] The release states "We reject this suggestion (that the SEC should not consider cases over an organized exchange because there are no 'face to face transactions'). It would be anomalous indeed if the protection afforded by antifraud provisions were withdrawn from transactions effected on exchanges, primary markets for securities transactions."[63] The release also tackled the subject of information whose release would be harmful to the firm. "If, on the other hand, disclosure prior to effecting a purchase or sale would be improper or unrealistic under the circumstances, we believe the alternative is to forego the transaction."[64]

---

[59] *Supra* note 40.
[60] *Ibid.*

Since transactions data are available only for the corporate insiders covered under the old antifraud provisions, the extensiveness of liability to tippees is not relevant to the study. However, the clarification of the provisions of the Act is relevant, as this may have deterred insider trading.

Appendix 3

Texas Gulf Sulphur case

■ On April 19, 1965, the Securities and Exchange Commission filed suit against the Texas Gulf Sulphur Company and a few high-level employees of the company. The employees were alleged to have traded on secret information concerning the possibility of mineral deposits on land in Ontario owned by the company. The ensuing case may well have been the most important one concerning the trading of insiders. Both the initial SEC indictment and the subsequent United States Court of Appeals decision expanded and clarified the legal restrictions placed on insiders. Because of this, the present study treats the indictment and subsequent decision as separate events.

In 1959, Texas Gulf Sulphur commenced a drilling operation in Ontario that indicated the possibility of huge mineral deposits. Numerous ruses, including "fake drill cores and the removal of the drilling rig to another location,"[65] were employed to conceal this information. On April 12, 1964, the company issued a false press release indicating that there was no conclusive evidence that mineral deposits had been discovered. The company reversed its position on April 16 by issuing a statement revealing that mineral deposits had previously been found.[66]

Since November 1963, Texas Gulf Sulphur employees were speculating in their company's security. The insiders enjoined as defendants purchased a total of 9,100 shares of stock between November 1963 and March 1964. In addition, insiders transferred information concerning the huge mineral deposits to others (commonly referred to as tippees). The tippees mentioned in the case purchased 12,100 shares of stock and purchased calls on 14,100 shares of stock. On November 10, 1963, the price of a share of Texas Gulf Sulphur stock was 17¾, while in April 1964, the price had moved to around 30. The stock price continued to rise and reached 71 within a year.[67]

---

[61] Manne [38], p. 39.

[62] For a full discussion of the special facts rule, see Loss [35], p. 825.

[63] *Supra* note 43.

[64] *Ibid.*

[65] Manne [38], p. 41.

[66] *Ibid.*

The action sought to compel the defendants to offer restitution to the individuals from whom the corporate insiders or tippees purchased stock or calls. The action also sought to cancel certain stock options and to require the return of all profits obtained from the exercise of other options. Furthermore, it was pleaded that Texas Gulf Sulphur Company be enjoined from issuing misleading press releases.[68]

The Texas Gulf Sulphur indictment has numerous implications for security law. Perhaps the most relevant is that "the Commission is seeking judicial approval for the rule it announced in Cady, Roberts and Co., namely, that corporate officers, directors, and controlling shareholders, and anyone else who had access to information intended to be available only for a corporate purpose may not take advantage of such material information by trading in the issuer's securities, while such information is unavailable to the persons with whom they are dealing."[69] Thus, this case action can be expected to deter insiders from speculating on inside information.

On August 19, 1966, a decision was handed down[70] declaring that all trading prior to April 9, 1964, was not unlawful because information concerning the mineral deposits was not considered to be material at that time. However, trading after April 9 was declared to be in violation of the law. The decision also stated that the press release was not illegal.

The decision has been appealed[71] and private suits have been instigated against Texas Gulf Sulphur.[72] However, these later actions have changed little the results reached in the first case. Thus, this study examined only the properties of insider trading before and after the date of the filing of the suit (April 19, 1965) and the properties of insider trading before and after the date of the first decision (August 19, 1966).

References

1. BALL, R. "Risk, Return, and Disequilibrium: An Application to Changes in Accounting Techniques." *Journal of Finance*, Vol. 27, No. 2 (May 1972), pp. 343–354.
2. ——— AND BROWN, P. "An Empirical Evaluation of Accounting In-

---

[67] *Ibid.*, pp. 41–43.

[68] Jennings and Marsh [31], p. 141.

[69] *Ibid.*, pp. 142–143.

[70] SEC v. Texas Gulf Sulphur Co., 401 F. Supp. 262 (S.D.N.Y. 1966).

[71] SEC v. Texas Gulf Sulphur Co., 401 F.2d 833 (2d Cir. 1968); SEC v. Texas Gulf Sulphur, 312 F. Supp. 77 (S.D.N.Y. 1970).

[72] In particular see Reynolds v. Texas Gulf Sulphur, 309 F. Supp. 548 (D. C. Utah 1970).

come Numbers." *Journal of Accounting Research,* Vol. 6, No. 2 (Autumn 1968), pp. 159–178.

3. *Bank & Quotation Record,* 1960–1970. William B. Dana Co.
4. Black, F. "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business,* Vol. 45, No. 3 (July 1972), pp. 444–454.
5. ———, Jensen, M. C., and Scholes, M. S. "The Capital Asset Pricing Model: Some Empirical Tests," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [30].
6. Black, F. and Scholes, M. S. "Dividend Yields and Common Stock Returns: A New Methodology." Working Paper No. 488–70, Sloan School of Management, Massachusetts Institute of Technology, September 1970.
7. Brownlee, K. A. *Statistical Methods in Science and Engineering.* New York: John Wiley, Inc., 1965.
8. Comment. "Class Action and 10b–5." *University of Chicago Law Review,* Vol. 38, No. 2 (Winter 1971), pp. 337–371.
9. Comment. "SEC Action Against Fraudulent Purchases of Securities." *Harvard Law Review,* Vol. 60, No. 1 (1946), pp. 769–780.
10. Comment. "Deterrence of Tippee Trading under Rule 10b-5." *University of Chicago Law Review,* Vol. 38, No. 2 (Winter 1971), p. 374.
11. Cootner, P. H., ed. *The Random Character of Stock Market Prices.* Cambridge: MIT Press, 1969.
12. Demsetz, H. "Information and Efficiency: Another Viewpoint." *Journal of Law and Economics,* Vol. 12, No. 1 (April 1969), pp. 1–22.
13. ———. "Perfect Competition, Regulation, and the Stock Market," in Henry A. Manne, ed., *Economic Policy and the Regulation of Corporate Securities.* Washington, D. C.: American Enterprise Institute, 1969.
14. Driscoll, T. E. "Some Aspects of Corporate Insider Stock Holdings and Trading under Section 16b of Securities and Exchange Act of 1934." Unpublished MBA Thesis, University of Pennsylvania, Philadelphia, Pennsylvania, 1963.
15. Fama, E. F. "Behavior of Stock Market Prices." *Journal of Business,* Vol. 38, No. 1 (January 1965), pp. 34–105.
16. ———. "Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance,* Vol. 25, No. 2 (May 1970), pp. 383–417.
17. ———, Fisher, L., Jensen, M. C., and Roll, R. "The Adjustment of Stock Prices to New Information." *International Economic Review,* Vol. 10, No. 1 (February 1969), pp. 1–21.
18. Fama, E. F. and Laffer, A. "Information and Capital Markets." *Journal of Business,* Vol. 44, No. 3 (July 1971), pp. 289–298.
19. Fama, E. F. and MacBeth, J. "Risk, Return and Equilibrium: Some Empirical Tests." Unpublished mimeograph, University of Chicago, Graduate School of Business, 1971.
20. Fama, E. F. and Miller, M. *The Theory of Finance.* New York: Holt, Rinehart and Winston, Inc., 1972.
21. Fisher, L. "Outcomes for Random Investments in Common Stocks Listed on the New York Stock Exchange." *Journal of Business,* Vol. 38, No. 2 (April 1965), pp. 149–161.
22. ———. "Some New Stock Market Indexes." *Journal of Business,* Vol. 39, No. 1, Supplement (January 1966), pp. 191–225.
23. Friend, I. "Discussion" in *Economic Policy and the Regulation of Corporate Securities.* Washington, D. C. American Enterprise Institute, 1969.
24. Glass, G. A. "Extensive Insider Accumulation as an Indicator of Near Term Stock Price Performance." Unpublished Ph.D. dissertation, Ohio State University, 1966.
25. Hirshleifer, J. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review,* Vol. 61, No. 4 (September 1971), pp. 561–574.
26. Ibbotson, R. "The Price Performance of New Issues of Common Stock." Unpublished paper, University of Chicago, 1973.
27. Jaffe, J. F. "Special Information, Security Regulation and Insider Trading." Unpublished Ph.D. dissertation, University of Chicago, 1972.

28. Jensen, M. C. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance*, Vol. 23, No. 2 (May 1968), pp. 389–416.
29. ———. "Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios." *Journal of Business*, Vol. 42, No. 2 (April 1969), pp. 167–247.
30. ———, ed. *Studies in the Theory of Capital Markets*. New York: Praeger Publishers, 1972.
31. Jennings, R. W. and Marsh, H. *Security Regulation: Cases and Materials—1966 Supplement*. Brooklyn: Foundation Press, 1966.
32. King, B. F. "Market and Industry Factors in Stock Price Behavior." *Journal of Business*, Vol. 39, Part II (January 1966), pp. 139–190.
33. Lintner, J. "Security Prices, Risk, and Maximal Gains from Diversification." *Journal of Finance*, Vol. 20, No. 5 (December 1965), pp. 587–615.
34. Lorie, J. H. and Neiderhoffer, V. "Predictive and Statistical Properties of Insider Trading." *Journal of Law and Economics*, Vol. 11, No. 1 (April 1968), pp. 35–51.
35. Loss, L. *Security Regulation*. Boston: Little, Brown and Co., 1951.
36. Mandelker, G. "The Effect of Mergers on the Value of the Firm." Unpublished manuscript, University of Chicago, 1973.
37. Manne, H. A. *Economic Policy and the Regulation of Corporate Securities*. Washington, D. C.: American Enterprise Institute, 1969.
38. ———. *Insider Trading and the Stock Market*. New York: Free Press, 1966.
39. Markowitz, H. *Portfolio Selection: Efficient Diversification of Investments*. New York: John Wiley, 1959.
40. Marshall, J. "Private Incentives and Public Information." Unpublished manuscript, Vanderbilt University, December, 1971.
41. Meyers, S. L. "A Re-examination of Market and Industry Factors in Stock Price Behavior." *Journal of Finance*, Vol. 28, No. 3 (June 1973), pp. 695–705.
42. Rogoff, D. L. "The Forecasting Properties of Insider Transactions." Unpublished DBA thesis, Michigan State University, 1964.
43. Scholes, M. S. "A Test of the Competitive Hypothesis: The Market for New Issues and Secondary Offerings." Unpublished Ph.D. dissertation, University of Chicago, 1969.
44. Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, Vol. 19, No. 4 (September 1964), pp. 425–442.
45. ———. "A Simplified Model for Portfolio Analysis." *Management Science*, Vol. 9, No. 2 (January 1963), pp. 277–293.
46. United States Securities and Exchange Commission. *Official Summary of Security Transactions and Holdings*.
47. Wu, H. K. "Corporate Insider Trading Profitability and Stock Price Movement." Unpublished Ph.D. dissertation, University of Pennsylvania, 1963.

SYMPOSIUM ON THE OPTIMALITY OF COMPETITIVE CAPITAL MARKETS

125 Production theory and the stock market
H. E. LELAND

145 The optimality of a competitive stock market
R. C. MERTON and
M. G. SUBRAHMANYAM

171 On the theory of the firm in an economy with incomplete markets
S. EKERN and R. WILSON

181 A note on unanimity of stockholders' preferences among alternative production plans: a reformulation of the Ekern-Wilson model
R. RADNER

Production theory and the stock market

Hayne E. Leland

Assistant Professor of Economics
Stanford University

Traditional economic models separate firms' production decisions from equilibrium in stock markets. In this paper, we develop an integrated model of production in the presence of capital asset market equilibrium. Our theory indicates that, in a stochastic environment, production and financial variables are inextricably interrelated. Following the financial equilibrium models of Sharpe, Lintner, and Mossin, we assume that profits and therefore portfolio returns are random. But stockholders can alter their distributions of returns by altering firms' production decisions as well as by altering their portfolios. The key to the analysis is a "unanimity theorem," which shows that in many environments stockholders will agree on optimal output decisions, despite their different expectations and attitudes towards risk. We develop equilibrium conditions which must be satisfied by production decisions. Profit maximization is indeed optimal for a firm whose profits are riskless. But risky firms' outputs depend on financial as well as cost variables, and the equilibrium conditions lead to a theory of production under uncertainty which replaces the now-vacuous notion of profit maximization. We further show that the output decisions will be Pareto optimal for stockholders, and that these decisions maximize market value only in a "purely competitive" world. Our results provide a synthesis of the conflicting conclusions of Diamond, Stiglitz, and Wilson on the optimality of stock markets.

1. Introduction

■ Theories of production and of finance have developed along remarkably independent paths. "The Stock Market" is not an

---

Hayne E. Leland received the A.B. from Harvard (1963), the M.Sc. (Econ.) from the London School of Economics (1965), and the Ph.D. from Harvard (1968). His current research deals with the economics of uncertainty, the extraction of natural resources, and the behavior of the regulated company.

This work was supported by National Science Foundation Grant GS-2874-A1 to the Institute for Mathematical Studies in the Social Sciences, Stanford University and by grants from the National Science Foundation and the Ford Foundation at Yale University. The author expresses his gratitude to Karl Shell, Robert Wilson, William Sharpe, and Joseph Stiglitz, who not only made valuable suggestions, but listened with remarkable patience to earlier versions of this work. Peter Diamond, William Brainard, and John Lintner also provided valuable insights. An earlier version of this paper appeared as "Stockholders' Interests, Value Maximization, and the Optimality of Stock Markets," Working Paper No. 20, Institute for Mathematical Studies in the Social Sciences, Stanford University.

essential element in neoclassical production models. And until recently, production has not appeared in models of capital asset market equilibrium.[1] Neoclassical production models have not needed stock markets because profit maximization provides a complete description of firm behavior under certainty. Stockholders are invoked to justify profit maximization, but thereupon are hastily retired from the scene. Indeed, the existence of the stock market has been somewhat embarrassing, since traditional economic models are complete without it.

The neoclassical approach breaks down, however, in the presence of uncertainty. The assumption of profit maximization no longer ranks alternative decisions, since profits are not uniquely determined by firms' actions. Production theory, and the welfare propositions constructed thereupon, must be reconstructed when the economic environment is stochastic.[2]

Extensions of traditional theory have been suggested to explain production decisions under uncertainty. The models of Sandmo, Lintner, and Leland[3] suggest that firms facing random prices act to maximize expected utility of profit. But in an environment in which owners possess differing attitudes towards risk as well as differing expectations, whose utility function and subjective probabilities are to be used?

Models of financial equilibrium, in contrast, have explicitly recognized the random character of profits, and the ownership of firms by diverse stockholders. But while the pioneering capital asset models of Sharpe, Lintner, and Mossin[4] have helped explain optimal portfolios and relative values of risky assets, they have shed no light on production, since these decisions are presumed fixed.

Not until Diamond's notable paper[5] were production decisions examined in the context of stock market equilibrium. Subsequent papers by Stiglitz, Jensen and Long, and Fama[6] have questioned the Pareto optimality of stock markets in the allocation of investment. All those studies have assumed that firms maximize market value. Yet Wilson,[7] using a variant of Stiglitz' model, shows that *every* stockholder would recommend production decisions which do *not* maximize market value. Wilson concludes that what is to be questioned is not the optimality of stock markets, but rather the presumption of value maximization by firms.[8] See Figure 1 for individual demand functions for shares under different policies.

---

[1] Recent contributions which examine production (or investment) include Diamond [5], Stiglitz [17], Jensen and Long [9], and Fama [8]. These studies are discussed below. Stigum [19] considers existence of general equilibria under uncertainty, and Dreze [6] develops a decentralized "tatonnement" process in a model quite similar to ours.

[2] An important exception to this is the case where separate contracts can be made for each state of nature: see Arrow [1]. (It seems clear that the actual environment does not provide such a complete set of markets.) See also Debreu [4].

[3] In [15], [11], and [10], respectively.

[4] In [16], [12], and [14], respectively.

[5] In [5].

[6] In [17], [9], and [8], respectively.

[7] In [20].

[8] Value maximization as the firm's goal has been defended on two

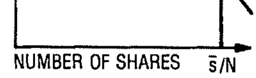
We shall build on an element underlying both finance and production theories: firms should operate "in the stockholders' interests." Production theories have interpreted it to imply profit maximization. Financial studies have interpreted it to imply value maximization. But neither approach has rigorously justified its interpretation, and Wilson's work seriously questions the validity of value maximization.

We develop below a model with production and with a stock market. Investors have diverse expectations and attitudes towards risk, and choose optimal portfolios given (equilibrium) market values. Firms face random prices for their products and must select outputs. These outputs are assumed to be selected in the stockholders' interests. A key to our analysis is the proof that, in competitive stochastic environments, outputs selected by firms in equilibrium will be in *all* stockholders' interests—regardless of their expectations or willingness to bear risk. The chosen outputs are Pareto optimal for stockholders, but in general will not maximize market value. We can show, however, that in certain "purely competitive" environments, firms acting *as if* they maximized market value will make choices which coincide with stockholders' interests. The contrasting results of Diamond, Stiglitz, and Wilson[9] are readily explicable by our theory.

Outputs chosen by firms will depend upon both production (costs) *and* financial (market value) variables. *Thus, our theory indicates that the stock market and production are inextricably connected under uncertainty.* For competitive firms facing nonrandom prices, stockholders will choose profit maximizing policies

FIGURE 1

INDIVIDUAL DEMAND FUNCTIONS FOR SHARES UNDER DIFFERENT POLICIES



$\bar{s}$ = FIXED NUMBER OF SHARES OUTSTANDING

N= NUMBER OF (IDENTICAL) INVESTORS

p^1= PRICE PER SHARE GIVEN POLICY 1

p^2= PRICE PER SHARE GIVEN POLICY 2

---

grounds: (1) a higher value implies all stockholders are better off; and (2) if policies are adopted which have a lower market value, entrepreneurs can (and will) buy the firm, change policies, and sell out at a sure profit. In fact, neither of these arguments is in general correct. For simplicity, it is useful to think of a world where there are the only two policies, and all investors are identical. In Figure 1, we have shown the individual demand curves of a typical investor for returns associated with the two policies. The identical investor assumption implies that each investor will purchase $\bar{s}/N$ shares, where $\bar{s}$ is the number of shares outstanding and N is the number of investors.

Note that policy 2 yields a higher equilibrium market value for the firm. But (using a consumer surplus argument) all investors would rather the firm choose policy 1. Hence, argument (1) is incorrect. Furthermore, to buy all shares of the firm, an entrepreneur would have to pay the equivalent of an "all or none" offer—again, the area *under* the curve D^1, which is greater than the amount he would receive after changing to policy 2, the area under D^2. Thus, argument (2) also is incorrect. In a recent paper, John Long [13] also argues that value maximization need not maximize welfare. Long restricts his attention to a mean-variance model.

If firms' returns are perfectly correlated, their shares are perfect substitutes, and the demand curve for a firm's shares is perfectly elastic. In this case, higher market values and preferred polices will coincide—see Theorem 5 below.

Although Stiglitz' published paper [17] treats only firms which maximize market value, his unpublished manuscript [18] clearly indicates his reservations about the appropriateness of value maximization. In the manuscript, he develops an analysis which parallels several of the ideals in Section 5, and indicates that different notions of equilibrium can give rise to different optimal strategies for stockholders.

[9] In [5], [17], and [20], respectively.

$(p = MC)$. For firms facing random prices, stockholders select outputs which behave like certainty outputs in some ways, but differently in others.

Our results, hopefully, point the way to further integration of the theories of production and finance, and to the reconstruction of value and welfare theory in stochastic environments. While we do not undertake empirical testing of the theory of this paper, it might be noted that our results do not require estimates of subjective probabilities for testing. In contrast with most other economic models under uncertainty, we need know only equilibrium market values, rather than the subjective probabilities which they reflect.

2. Firms facing random prices

■ We consider an economy with firms indexed $j = 0, \ldots, N$. A firm has a profit function relating profit π^j to the vector of decisions of the firm d^j and the unknown state of nature θ:

$$\pi^j = \pi^j(d^j, \theta).^{10} \tag{1}$$

Randomness could enter from any or all of several factors, including uncertain demand or uncertain technology. In the present section, we limit ourselves to firms similar to the competitive models of Sandmo and Baron,[11] where selling price is random but cost functions are known. In this environment,

$$\pi^j = p^j(\theta)q^j - C^j(q^j), \quad j = 0, \ldots, N, \tag{2}$$

where q^j is the (scalar) output decision of the jth firm. Each firm owns certain resources essential to production, such as capital, "goodwill," or technical expertise. $C^j(q^j)$ reflects the costs incurred by the hiring of further factors of production, and as such could be viewed as a "variable cost" function. We assume there exists a firm which faces a certain price:[12]

$$\pi^0 = \bar{p}^0 q^0 - C^0(q^0). \tag{3}$$

Subsequent sections consider noncompetitive environments, multiple decision variables, and technological uncertainty.

Firms issue shares, which represent a claim to profits, such that the ownership of a fraction s_i^j of the shares of firm j entitles the ith stockholder to a return $s_i^j\pi^j$. We denote the current total market value of the jth firm's shares by V^j, and define $V = (V^0, \ldots, V^N)$.[13]

---

[10] To formalize our notion of "states of nature," we presume that θ is an element in a measurable space (Θ, Σ), where Σ is a σ-field of subsets of Θ ("events"). If D is a feasible set of decisions, then π^j is a measurable function from $D \times \Theta$ into R^1, for all j. The specification of a probability measure over Σ will then generate a joint probability distribution of profits of firms, conditional on d.

[11] In [15] and [3].

[12] Alternatively, we could have assumed the existence of a riskless bond paying a fixed coupon rate.

[13] Note our total market value is the market value of equities. For simplicity, we ignore possible debt financing. But one can show that, if debt is not issued beyond the point of possible bankruptcy (as perceived by any investor), the Modigliani-Miller theorem survives: see note 27.

3. Investors

■ We assume there are M investors, indexed $i = 1, \ldots, M$. Investors are described by three attributes:

(1) $\bar{s}_i = (\bar{s}_i^0, \ldots, \bar{s}_i^N)$, the current portfolio measured in fractions of each firm's stock, owned by investor i. We assume $\Sigma_i \bar{s}_i^j = 1$, $j = 0, \ldots, N$.

(2) $U_i(R_i, \theta)$ i's utility function, unique to a positive linear transformation, which gives the utility of return R_i in state θ.

We assume $U'_i(R_i, \theta) \equiv \dfrac{\partial U_i(R_i, \theta)}{\partial R_i} > 0$ (nonsatiation); $U''_i(R_i, \theta) \equiv \dfrac{\partial^2 U_i(R_i, \theta)}{\partial R_i^2} < 0$ (concavity in R_i).

(3) μ_i, the ith investor's subjective probability measure over the states of nature.[14] E_i is the expectation with respect to μ_i.

Investors are assumed to rank alternatives by expected utility.[15] Given arbitrary market values $V = (V^0, \ldots, V^N)$ of firms' stocks, we can define budget sets $B(V, \bar{s}_i)$ for each investor:

$$B(V, \bar{s}_i) = \left\{ s_i \in R^{N+1} \,\middle|\, \sum_{j=0}^{N} V^j (s_i^j - \bar{s}_i^j) \leqq 0 \right\}. \tag{4}$$

Clearly, $B(V, \bar{s}_i)$ gives the set of portfolios which an investor could afford when market values are V and his current portfolio is $\bar{s}_i$. Nonsatiation implies that *optimal* portfolios will be restricted to the subset of $B(V, \bar{s}_i)$ such that

$$\sum_{j=0}^{N} V^j (s_i^j - \bar{s}_i^j) = 0, \quad \text{for all } i. \tag{5}$$

Following Diamond and Stiglitz,[16] we define R_i, the return to a portfolio s_i, as

$$R_i = \sum_{j=0}^{N} \pi^j s_i^j, \tag{6}$$

or substituting for s_i^0 from (5),

$$R_i = r \sum_{j=0}^{N} V^j \bar{s}_i^j + \sum_{j=1}^{N} (\pi^j - r V^j) s_i^j, \tag{7}$$

where $r \equiv \pi^0(q^0)/V^0$, the return on the riskless asset.[17]

---

[14] More exactly, μ_i is a probability measure on the measurable space (Θ, Σ). Thus the specification of different probability measures for different investors implies each investor has his own subjective probability distribution for the profits of firms, conditional upon their decisions.

[15] Since the utility function in attribute (2) is state dependent, we are dealing with a more general form of expected utility than is typical. Peter Diamond has pointed out that all theorems are valid in the more general case where (assuming there are a finite number of states of nature) choice between decisions can be ranked by a nonsatiated, quasi-concave function

$$U[R_i^1(d), \ldots, R_i^S(d)],$$

where $R_i^S(d)$ is the return to investor i in state of nature s, given decisions d.

[16] In [5] and [17], respectively.

[17] The assumption that returns to holding shares depend only on firms'

4. Financial equilibrium

■ Portfolios $\hat{s}_i = (\hat{s}_i^0, \ldots, \hat{s}_i^N)$, $i = 1, \ldots, M$, and market values $\hat{V} = (\hat{V}^0, \ldots, \hat{V}^N)$ constitute a *financial equilibrium* relative to production decisions $q = (q^0, \ldots, q^N)$ if:

(i) For each investor, i, $\hat{s}_i$ is optimal in $B(\hat{V}, \bar{s}_i)$, given q. That is, $\hat{s}_i$ maximizes $E_i[U_i(R_i, \theta)]$ for $s_i \in B(\hat{V}, \bar{s}_i)$, given q.

(ii) $\hat{V}$ equates supply and demand in each security's market, given q. That is, $\Sigma_i \hat{s}_i^j = 1$, $j = 0, \ldots, N$.

Financial equilibrium is a familiar concept. Condition (i) states that each consumer is in equilibrium given market values $\hat{V}$. Condition (ii) states that values $\hat{V}$ are equilibrium market values. Given fixed q and further assumptions about subjective probability functions and preferences, Sharpe[18] and others have examined properties of equilibrium portfolios and market values.

In our model, it is clear that equilibrium portfolios and market values depend upon q. This is because a change in q will change the distributions of profits perceived by each investor. Let us denote the *financial equilibrium correspondence* by

$$[\hat{s}_1(q), \ldots, \hat{s}_M(q);\ \hat{V}(q)].$$ [19]

We assume that, for each q in a relevant domain, a financial equilibrium exists. Later, we shall assume $\hat{s}_i(q)$ and $\hat{V}(q)$ are differentiable functions (as they will be in a mean-variance framework).

Conditions (i) of financial equilibrium imply $\hat{s}_i$ is optimal, given q, for all i. $(\hat{s}_i^1, \ldots, \hat{s}_i^N)$ will maximize (with no constraints) $E_i[U_i(R_i, \theta)]$ when s_i^0 is eliminated by (5) and R_i is given by (7). Necessary first-order maximizing conditions are

$$E_i\{U'_i(R_i, \theta)[\pi^j - r\hat{V}^j]\} = 0;\ j = 1, \ldots, N,\ i = 1, \ldots, M. \tag{8}$$

When profits can be written in the competitive form (2), we may rewrite (8) as

$$E_i\{U'_i(R_i, \theta)[p^j(\theta)q^j - C^j(q^j) - r\hat{V}^j]\} = 0. \tag{9}$$

---

profits is adopted by Diamond [5], Stiglitz [17], and others. This assumption is reasonable in a stationary world with intertemporal independence, or in a world in which the movement of stock prices is closely dependent upon profits, as seems reasonable in the long run. In the short run, however, speculators may have (probabilistic) expectations of returns which bear little or no connection to firms' decisions. If speculators perceive no connection, they can be included in our model without altering the results which follow. The diligent reader also can ascertain that a two-period consumption model will not alter the nature of our results. An alternative formulation to (6), in which investors rank portfolios according to end-period wealth, is considered in Section 9.

[18] In [16].

[19] Of course, the financial equilibrium also depends on the current portfolios $\bar{s}_i$, but this dependence is suppressed since we shall not consider variations in $\bar{s}_i$. Also note that the financial equilibrium correspondence gives optimal portfolios and equilibrium market values if q is regarded as fixed—i.e. in equilibrium. If all investors know q will change to some $\bar{q}$, and values will change to $\hat{V}(\bar{q})$, then $\hat{V}(q)$ may no longer describe the actual market values at q.

Conditions (ii) of financial equilibrium imply

$$\sum_i \hat{s}_i^j = 1, \quad j = 0, \dots, N. \tag{10}$$

With (5), (8) or (9), and (10), $(N+1)(M+1)$ equations are formed in the $(N+1)(M+1)$ unknowns $\hat{s}_i^j$ and $\hat{V}^j$, $i = 1, \dots, M$; $j = 0, \dots, N$. Because these equations are homogeneous of degree zero in $\hat{V}$, only relative prices are determined, and we normalize prices by choosing

$$\hat{V}^0 = 1, \quad \text{for all } q.^{20} \tag{11}$$

This implies that the equilibrium correspondence of market values takes the form $\hat{V}(q) = [1, \hat{V}^1(q), \dots, \hat{V}^N(q)]$, and that

$$r = \frac{\pi^0(q^0)}{\hat{V}^0(q)} = \pi^0(q^0). \tag{12}$$

5. Equilibrium production decisions

■ A vector of outputs $\hat{q} = (\hat{q}^0, \dots, \hat{q}^N)$ is termed a *production equilibrium* if outputs $\hat{q}$ are "in the stockholders' interests," when current portfolios and market values are a financial equilibrium relative to $\hat{q}$.

The requirement that current portfolios and market values be a financial equilibrium given $\hat{q}$, i.e. that $[\bar{s}_i = \hat{s}_i(q); \bar{V} = \hat{V}(q)]$, implies there is no impetus for portfolio or market value changes if $\hat{q}$ remains fixed. And the requirement that $\hat{q}$ be "in the stockholders' interests" implies that $\hat{q}$ will not change, given current portfolios and market values.

We have purposefully left imprecise the meaning of outputs being "in the stockholders' interests." Any number of voting rules might actually be used, as long as they lead to some set of outputs being chosen. But *whatever the rule* for determining output, we presume there cannot be some $\bar{q}^k \neq \hat{q}^k$, and associated (different) financial equilibrium, such that *all stockholders* of k are better off at $\bar{q}^k$ than at $\hat{q}^k$. Clearly, if such a $\bar{q}^k$ exists, and stockholders influence the firm (or the manager is a stockholder), then the firm will change output from $\hat{q}^k$, and $\hat{q}^k$ cannot represent a production equilibrium.

To examine the implications of production equilibrium, we first prove a theorem on stockholder unanimity.

Theorem 1: *Unanimity theorem.* In the competitive random environment (2), assume current portfolios and market values are in financial equilibrium given current production decisions $q = (q^0, \dots, q^N)$. Then $\dfrac{\partial E_i[U_i(R_i, \theta)]}{\partial q^k}$ has the same sign for all stockholders. That is, stockholders will unanimously vote for (or against) small changes in output.

Proof:

(a) For the riskless firm:

---

[20] If we start with a two-period consumption model, this normalization will not be possible. Nonetheless, all our results will follow.

$$\frac{\partial E_i[U_i(R_i, \theta)]}{\partial q^0} = E_i\left[U'_i(R_i, \theta)\left(\frac{\partial R_i}{\partial q^0}\right)\right], \tag{13}$$

where R_i is given by (7), r by (12), and $\bar{s}_i = \hat{s}_i(q)$. Since $\hat{V}(q)$, r and $\hat{s}_i$ will change as q^0 changes:

$$\begin{aligned}
\frac{\partial E_i[U_i(R_i, \theta]}{\partial q^0} &= E_i\left\{U'_i(R_i, \theta)\left\{\left[[\bar{s}_i^0 + \sum_{j=1}^{N}[\hat{V}^j(\bar{s}_i^j - \hat{s}_i^j)]\right]\frac{dr}{dq^0}\right.\right. \\
&\quad \left.\left. + r\sum_{j=1}^{N}\left[(\bar{s}_i^j - \hat{s}_i^j)\frac{\partial \hat{V}^j}{\partial q^0}\right] + \sum_{j=1}^{N}\left[(\pi^j - r\hat{V}^j)\frac{\partial \hat{s}_i^j}{\partial q^0}\right]\right\}\right\} \\
&= E_i\left[U'_i(R_i, \theta)\left(\frac{dr}{dq^0}\right)\bar{s}_i^0\right], \text{ using (8) and } \bar{s}_i = \hat{s}_i; \\
&= E_i\left[U'_i(R_i, \theta)\frac{d\pi^0}{dq^0}\right]\bar{s}_i^0, \text{ using (12);} \\
&= E_i[U'_i(R_i, \theta)]\bar{s}_i^0[p^0 - MC(q^0)], \quad \text{using (3),}
\end{aligned} \tag{14}$$

where $MC^j(q^j) \equiv \dfrac{dC^j(q^j)}{dq^j}$, $j = 0, \ldots, N$. Now $E_i[U'_i(R_i, \theta)]\bar{s}_i^0$ is positive (by nonsatiation) for all investors with $\hat{s}_i^0 = \bar{s}_i^0 > 0$; that is, *for all stockholders.* Therefore (14) has the sign of $[p^0 - MC^0(q^0)]$ for every stockholder.

(b) For risky firms:

For arbitrary k, $1 \leqq k \leqq N$,

$$\begin{aligned}
\frac{\partial E_i[U_i(R_i, \theta)]}{\partial q^k} &= E_i\left[U'_i(R_i, \theta)\left(\frac{\partial R_i}{\partial q^k}\right)\right] \\
&= E_i\left\{U'_i(R_i, \theta)\left[r\sum_{j=1}^{N}\left[(\bar{s}_i^j - \hat{s}_i^j)\frac{\partial \hat{V}^j}{\partial q^k}\right]\right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{N}\left[(\pi^j - r\hat{V}^j)\frac{\partial \hat{s}_i^j}{\partial q^k}\right] + \frac{d\pi^k}{dq^k}\hat{s}_i^k\right]\right\} \\
&= E_i\left\{U'_i(R_i, \theta)\left[\frac{d\pi^k}{dq^k}\bar{s}_i^k\right]\right\}, \text{ using (8) and } \bar{s}_i = \hat{s}_i \\
&= E_i\{U'_i(R_i, \theta)[p^k(\theta) - MC^k(q^k)]\bar{s}_i^k\}, \quad \text{using (2).}^{21}
\end{aligned} \tag{15}$$

Now from conditions (9), since $\hat{s}_i = \bar{s}_i$ is a financial equilibrium,

---

[21] Note that our conclusion follows for any possible $\dfrac{\partial \hat{V}^j}{\partial q^k}$. Indeed, different investors can have different opinions regarding $\dfrac{\partial \hat{V}^j}{\partial q^k}$ and our results will still follow, as long as all investors face the equilibrium price $\hat{V}^j(\hat{q})$.

$$E_i\{U'_i(R_i, \theta)[p^k(\theta)]\} = E_i\{U'_i(R_i, \theta)\}\left[\frac{C^k(q^k) + r\hat{V}^k}{q^k}\right]. \tag{16}$$

Substituting (16) into (15) gives

$$\frac{\partial E_i[U_i(R_i, \theta)]}{\partial q^k} = E_i[U'_i(R_i, \theta)]\bar{s}_i^k\left[\frac{C^k(q^k) + r\hat{V}^k}{q^k} - MC^k(q^k)\right]. \tag{17}$$

But as before, since $E_i[U'_i(R_i, \theta)]\bar{s}_i^k > 0$ for all stockholders of k, the sign of (17) will be *identical* for all stockholders, and will have the same sign as $\frac{C^k(q^k) + r\hat{V}^k}{q^k} - MC^k(q^k)$. This last expression, of course, depends only on observable variables: cost functions and current market values.

Theorem 1 leads immediately to conditions which will be satisfied by equilibrium outputs in a competitive environment.

Theorem 2: *Fundamental equilibrium conditions.* If $\hat{q} = (\hat{q}^0, \ldots, \hat{q}^N)$ is a *production equilibrium,* then

(a) For the riskless firm:

$$p^0 = MC^0(\hat{q}^0); \tag{18}$$

(b) For risky firms:

$$\frac{C^j(\hat{q}^j) + r\hat{V}^j(\hat{q})}{\hat{q}^j} = MC^j(\hat{q}^j), \quad j = 1, \ldots, N. \tag{19}$$

Proof: If conditions (18) and/or (19) are not satisfied at $\hat{q}$, then a small change in some q^k can occur after which *all* stockholders in firm k are better off, since $\frac{\partial E_i[U_i(R_i, \theta)]}{\partial q^k} > 0$ (or < 0) for all stockholders of k by Theorem 1. But this contradicts the assumption that $\hat{q}$ is a production equilibrium.[22]

---

[22] At the chosen outputs, (18) and (19) and appropriate second-order conditions imply that all stockholders find $\hat{q}$ a local optimum. I have not been able to show that $\hat{q}$ represents a global optimum for all stockholders, although it will be in the mean-variance framework of Section 10, when cost functions are convex.

An alternative proof of the equilibrium condition (19) can be based on an examination of arbitrage. Assume stockholder i (with $\bar{s}_i^k > 0$) could change the output of firm k by Δq^k. The resulting change in his return will be $[p^k(\theta) - MC^k(q^k)]\bar{s}_i^k\Delta q^k$. At the same time, assume he exchanges a fraction of his shares $\Delta s_i^k = \frac{\bar{s}_i^k\Delta q^k}{q^k}$ for $V^k\Delta s_i^k$ shares of the riskless asset, producing a change $\left[-p^k(\theta) + \frac{C^k(q^k) + r\hat{V}^k}{q^k}\right]\bar{s}_i^k\Delta q^k$. The net return of simultaneous changes is

$$\left[\frac{C^k(q^k) + r\hat{V}^k}{q^k} - MC^k(q^k)\right]\bar{s}_i^k\Delta q^k,$$

whose sign is independent of θ and of i. Thus, all stockholders would urge changes in q^k until $\hat{q}^k$ is achieved and (19) is satisfied. Even if no riskless asset is present, the arbitrage argument can be used when the distributions of marginal returns are spanned by the distributions of total returns: see Ekern and Wilson [7].

■ Equilibrium condition (18) confirms what every schoolboy knows: the competitive firm under certainty sets price equal to marginal cost, and therewith maximizes profits. Thus, our theory confirms that profit maximization is optimal from the point of view of stockholders if profits are nonstochastic.

Under uncertainty, we showed earlier that profit maximization and "$p = MC$" are meaningless concepts. Conditions (19) replace the familiar relation (18). We can interpret (19) as follows: a risky firm j should choose a $\hat{q}^j$ such that the "total" average cost equals marginal cost, where the total average cost is the sum of average variable costs $\frac{C^j(\hat{q}^j)}{\hat{q}^j}$ plus an *imputed* average capital cost $\frac{r\hat{V}^j(\hat{q})}{\hat{q}^j}$.[23] Thus stock market value of the firm becomes crucial in the firm's output decision—our theory indicates the inextricable relationship between production (cost) and financial (market value) variables.

If the firm's physical assets are transferable and marketed, the analysis can be extended to consider entry and exit of firms. If $\hat{V}^j$, the market value of the jth firm, exceeds the market value of its assets, entry will take place in a competitive environment, since any entrepreneur who takes $\hat{V}^j$ as given perceives a sure profit by purchasing physical assets and then selling stock (whose market value will exceed the cost of the physical assets). In long-run competitive equilibrium, the value of firms will equal the value of their physical assets.[24]

Note how, in long-run equilibrium with entry, the firm will be producing *at the bottom of its (long-run) average cost curve.* This is because in long-run equilibrium $r\hat{V}^j$ is the true opportunity cost of the firm's capital stock, since $\hat{V}^j$ will equal the actual value of the firm's capital. Conditions (19) now imply true average cost equals marginal cost at the equilibrium output, implying the firm under uncertainty will produce at minimum average cost. If expectations are identical, the results in Section 10 imply that expected price will exceed marginal cost in equilibrium. Thus, while with free entry firms are producing at minimum average cost, their expected profits will be positive, and total output of the industry will be less than if p^j equalled its expected value with certainty. Greater uncertainty, then, leads to fewer firms and higher expected prices in long-run equilibrium, though all firms will be producing at the minimum average cost.

---

[23] Note that $\frac{r\hat{V}(\hat{q})}{\hat{q}}$ is the average opportunity cost of the market value of the firm. If the firm is a monopoly, this market value may be considerably higher than the true average cost of the capital stock in that industry.

(Clearly, readers should resist the temptation to conclude that "average cost equals marginal cost implies zero profits." First, profits are random so "zero profits" is meaningless. Second, we showed above that our "average cost" is dependent on the stock market value of the firm, which may not be closely related to the true value of the firm's capital stock.)

[24] In our model, the signal for entry is not when "excess" profits are present, but rather when market value exceeds the cost of duplicating a firm. Clearly this is a much more satisfactory situation in the presence of uncertainty. Even if *expected* profit could be determined under uncertainty, it would still have to be "risk adjusted" in some manner to make the old criterion for entry operable.

Conditions (19) can be used to generate useful theorems on firm behavior.

Theorem 3. $\hat{V}^j \geqq 0$ if and only if $MC^j(\hat{q}^j) \geqq AC^j(\hat{q}^j)$, where $AC^j(\hat{q}^j) = \dfrac{C^j(\hat{q}^j)}{\hat{q}^j}$.

Proof: From (19),

$$\hat{V}^j(\hat{q}) = \frac{\hat{q}^j}{r}\,[MC^j(\hat{q}^j) - AC^j(\hat{q}^j)], \tag{20}$$

which is nonnegative if and only if $MC^j(\hat{q}^j) \geqq AC^j(\hat{q}^j)$.

In long-run equilibrium, we shall expect $\hat{V}^j \geqq 0$.[25] This implies that optimal output $\hat{q}^j$ will be chosen so $MC^j(\hat{q}^j) \geqq AC^j(\hat{q}^j)$, or in the region where marginal cost lies above average (variable) costs. Note the similarity with the traditional theory of the competitive firm under certainty, which says $p = MC(\hat{q})$ for $p \geqq AVC(\hat{q})$, where AVC is average variable cost. Therefore, $MC(\hat{q}) \geqq AVC(\hat{q})$, which is precisely the condition we derive for the uncertainty case! *Thus market values being nonnegative is the substitute for long-run profits being nonnegative under certainty.*

Theorem 4. Any change in expectations, attitudes towards risk, etc., which leads to a rise in the firm's market value, but does not affect costs, will lead to greater output by the firm.

Proof: If the $\hat{V}^j(q)$ function shifts upwards to $V^j_*(q)$, all investors will find

$$\frac{C^j(\hat{q}^j) + rV^j_*(\hat{q}^j)}{\hat{q}^j} - MC^j(\hat{q}^j) > 0,$$

which by (17) implies all stockholders will urge an increase in q^j until a new equilibrium (19) is reached at a greater $\hat{q}^j$.[26]

Theorem 4 is a replacement for the certainty statement that "a rise in price leads to a rise in output." Since price is not defined under uncertainty, we see that market value takes its place in this context.

Theorem 5. In general, an increase in fixed costs will lead to a change in output.

Proof: If there is no change in optimal output, then

$$\frac{C(\hat{q}) + \Delta F + r(\hat{V} + \Delta\hat{V})}{\hat{q}} = MC(\hat{q}) = \frac{C(\hat{q}) + r\hat{V}(\hat{q})}{\hat{q}}.$$

or $\Delta\hat{V}/\Delta F = -1/r$. In general, it can be shown that $\Delta\hat{V}/\Delta F =$

---

[25] Free disposal of securities will guarantee $\hat{V}^j \geqq 0$.

[26] For the theorem to hold when changes in V are large and when stockholders are permitted to suggest nonmarginal changes, we would further require $E_t[U_t(R_t, \theta)]$ to be globally concave in q. Local concavity is assured by the fact that $\hat{q}$ provides a local maximum in $E_t[U_t(R_t, \theta)]$, for all stockholders.

$-1/r$ if and only if each investor exhibits constant absolute risk aversion.[27]

7. Stockholders' directives and value maximization

■ Will the q^j chosen by stockholders also maximize the market value of the firm? In general, the answer is no. But under conditions which we argue are truly "competitive," the stockholders' policies will maximize *perceived* market value.

First, we demonstrate that equilibrium outputs $q^j = \hat{q}^j$ will not in general maximize market value. For simplification, we use the special case of a single risky firm in addition to the riskless firm, although we still suppose firms act as if prices are independent of their decisions. Differentiating (8) with respect to $\hat{q}^1$, the risky firm's output, and evaluating at $q^1 = \hat{q}^1$ gives

$$E_i\left\{U'_i(R_i, \theta)\left[\frac{d\pi^1}{dq^1} - r\frac{\partial \hat{V}^1}{\partial q^1}\right]\right\} + E_i\left\{U''_i(R_i, \theta)(\pi^1 - r\hat{V}^1)\left[(\pi^1 - r\hat{V}^1)\frac{\partial \hat{s}_i^1}{\partial q^1} + \frac{d\pi^1}{dq^1}\hat{s}_i^1\right]\right\} = 0 \tag{21}$$

or, using (15) and the fact that in equilibrium

$$E_i\left[U'_i(R_i, \theta)\frac{d\pi^1}{dq^1}\right] = 0,$$

$$-D_i^{-1}E_i[U'_i(R_i, \theta)]\left(r\frac{\partial \hat{V}^1}{\partial q^1}\right) + \frac{\partial \hat{s}_i^1}{\partial q^1} + D_i^{-1}\left[E_i\left\{U''_i(R_i, \theta)(\pi^1 - r\hat{V}^1)\frac{d\pi^1}{dq^1}\hat{s}_i^1\right\}\right] = 0,$$

where $D_i \equiv E_i\{U''_i(R_i, \theta)[\pi^1 - r\hat{V}^1]^2\} < 0$. Now summing (21) over individuals $i = 1, \ldots, M$ and noting $\sum_i \frac{\partial \hat{s}_i^1}{\partial q^1} = 0$ gives

$$\left.\frac{\partial \hat{V}^1}{\partial q^1}\right|_{q^1 = \hat{q}^1} = \frac{\sum_i D_i^{-1}E_i\left[U''_i(R_i, \theta)(\pi^1 - r\hat{V}^1)\frac{d\pi^1}{dq^1}\hat{s}_i^1\right]}{r\sum_i D_i^{-1}E_i[U'_i(R_i, \theta)]}. \tag{22}$$

---

[27] The proof relies on the fact that there are no wealth effects if an investor possesses a utility function exhibiting constant absolute risk aversion. Thus a rise in fixed costs with an offsetting change in market cost (by a factor $1/r$) will lead all investors to demand precisely the same percentage of the firm's stock as before. Whereas Theorem 5 might seem to indicate that optimal output depends on the debt-equity ratio, such is not the case when a firm's debt is risk free. Each investor can always generate the same distribution of returns, independent of the debt-equity ratio, by an appropriate holding of the firm's debt and equity. (This is not possible when there is simply an increase in fixed cost.) There is good reason to think, however, that firms will be motivated to issue debt beyond the "zero probability of bankruptcy" level. We hope to explore this topic in a future paper.

By nonsatiation and $D_i < 0$, the denominator is always negative. In general, the numerator will be nonzero, which explains why Stiglitz and others can show that value maximization does not lead to Pareto optimal choices of outputs (or "techniques").[28] If the firm chooses a $q_V^j \neq \hat{q}^j$, our model indicates *everyone* could be made better off by a small change in q^j away from q_V^j. We support Wilson's contention that what Stiglitz brings into question is not the Pareto optimality of the stock market, but rather the value maximization criterion.

There is, however, an environment in which firms, *thinking* they are maximizing market value, make decisions which coincide with the $\hat{q}^k$ that stockholders wish.

Theorem 6: *Value maximization in a competitive environment.* Assume that a firm operates in an industry in which firms face the same random price and possess identical strictly convex cost functions. Then the firm will maximize its market value *relative to other firms in the industry* by setting $q^k = \hat{q}^k$, the output which is unanimously supported by stockholders.

Proof: Consider firms k and l in industry A, facing common price $P^A(\theta)$ and having identical cost functions $C^A(\cdot)$. Assume $q^k = \hat{q}^k$, where q^k is the shareholders' desired policy which satisfies (19). Further assume $q^l \neq \hat{q}^k$. We show that in the financial equilibrium associated with those outputs,

$$\hat{V}^l < \hat{V}^k.$$

From (9), for all i we know

$$E_i\{U'_i(R_i, \theta)[p^A(\theta)\hat{q}^k - C^A(\hat{q}^k) - r\hat{V}^k]\} = 0 \qquad (23)$$
$$E_i\{U'_i(R_i, \theta)[p^A(\theta)q^l - C^A(q^l) - r\hat{V}^l]\} = 0,$$

which imply

$$\hat{V}^k > \hat{V}^l \quad \text{if and only if} \quad E_i\{U'_i(R_i, \theta)[p^A(\theta)\hat{q}^k - C^A(\hat{q}^k)]\}$$
$$> E_i\{U'_i(R_i, \theta)[p^A(\theta)q^l - C^A(q^l)]\}. \qquad (24)$$

From (15) and the requirement $\dfrac{\partial E_i[U_i(R_i, \theta)]}{\partial q^k} = 0,$

$$E_i[U'_i(R_i, \theta)p^A(\theta)] = E_i\{U'_i(R_i, \theta)[MC^A(\hat{q}^k)]\}.$$

Substituting this into (24) and noting $E_i[U'_i(R_i, \theta)] > 0$ gives

$$\hat{V}^k > \hat{V}^l \quad \text{if and only if} \quad MC^A(\hat{q}^k)\hat{q}^k - C^A(\hat{q}^k)$$
$$> MC^A(\hat{q}^k)q^l - C^A(q^l). \qquad (25)$$

The inequality clearly will hold if

$$\operatorname*{Max}_q[MC^A(\hat{q}^k)q - C^A(q)] = MC^A(\hat{q}^k)\hat{q}^k - C^A(\hat{q}^k), \quad (26)$$

and the maximum will be unique since $C^A(\cdot)$ is strictly convex.

---

[28] In the case of a single risky firm, $\dfrac{\partial \hat{V}'}{\partial q'}$ can be shown to be negative at $\hat{q}'$. The author thanks Robert Forsythe for providing a proof in personal correspondence.

Now a necessary and sufficient condition to maximize the LHS of (26) is that

$$\frac{d}{dq}[MC^A(\hat{q}^k)q - C^A(q)] = MC^A(\hat{q}^k) - MC^A(q) = 0, \tag{27}$$

or $q = \hat{q}^k$. The theorem then follows immediately from (25).

Theorem 6 is definitely in the spirit of "perfect competition." If a firm maximizes its relative value but considers others' values independent of its decisions, it thinks it is maximizing its own market value. Thus competitive firms, attempting to maximize value by maximizing relative value, will make output choices which coincide with stockholders' interests. This result justifies Diamond's assumption that competitive firms should maximize a market valuation function of the form he chooses. Note that Stiglitz, in excluding perfectly correlated firms, excludes our notion of a "perfectly competitive" environment.

A final question to be asked is whether a firm acts "as if" it maximized expected utility of profit. In general, the answer is no, but if all shareholders are alike (e.g., like "the manager"), and prices of different firms are independently distributed, then it can be shown that q^j is "as if" firm j were maximizing expected utility of profit, and the theorems of Baron, Sandmo,[29] and Leland are valid in this environment.

8. Pareto optimality of production equilibrium

■ We now examine the optimality properties, from the point of view of stockholders, of a production equilibrium $\hat{q}$ and its associated financial equilibrium.[30] Pareto optimality requires that if every person l except an arbitrary individual m has fixed utility $E_l[U_l(R_l, \theta)] = \overline{U}_l$, then the output and portfolio decisions must maximize $E_m[U_m(R_m, \theta)]$. Feasibility requires $\sum_{i=1}^{M} s_i^j = 1$ for all j, so we can express the portfolio of the mth individual as

$$s_m = \left(1 - \sum_{l \neq m} s_l^0, \ldots 1 - \sum_{l \neq m} s_l^N\right).$$

If a production equilibrium is to be Pareto optimal, $\hat{q}$ and $\hat{s}_i(\hat{q}), i = 1, \ldots, M$, must yield a stationary point to the Lagrangian

$$L = E_m[U_m(R_m, \theta)] + \sum_{l \neq m} \lambda_l \{E_l[U_l(R_l, \theta)] - \overline{U}_l\},$$

implying

$$\frac{\partial L}{\partial q^k} = E_m\left[U'_m(R_m, \theta)\frac{\partial \pi^k}{\partial q^k}\right]\hat{s}_m^k + \sum_{l \neq m} \lambda_l E_l\left[U'_l(R_l, \theta)\frac{\partial \pi^k}{\partial q^k}\right]\hat{s}_l^k = 0, \tag{28}$$

---

[29] In [3] and [15], respectively.

[30] The optimality of outputs in our model is with respect to the distribution of profits they generate. Since our model is not closed (prices are exogenous, given θ), we cannot examine optimality of outputs from the point of view of their consumption.

$$k = 0, \ldots, N;$$

$$\frac{\partial L}{\partial s_l^j} = -E_m[U'_m(R_m, \theta)\pi^j] + \lambda_l E_l[U'_l(R_l, \theta)]\pi^j = 0,$$

$$j = 0, \ldots, N; l \neq m. \tag{29}$$

Now (14), (15), and the equilibrium requirements (18) and (19) imply that (28) will be satisfied by the production equilibrium $\hat{q}$ and its associated financial equilibrium, for arbitrary λ_l's. Using conditions (8), we see that (29) will be satisfied when $\lambda_l = \frac{E_m[U'_m(R_m, \theta)\pi^j]}{E_l[U'_l(R_l, \theta)\pi^j]}$, $l \neq m$. These λ_l's, in conjunction with $\hat{q}$ and $\hat{s}_i(\hat{q})$, $i = 1, \ldots, M$, constitute a stationary point to the Lagrangian L. Thus, we conclude that production equilibrium satisfies the first-order necessary conditions for Pareto optimality. If $\hat{s}_i^j(\hat{q})$ is positive for all i and j, then $E_m[U_m(R_m, \theta)]$ will indeed be maximized for arbitrary m, and we conclude that production equilibrium is Pareto optimal from the point of view of the stockholders.[31]

9. Some generalizations: a preliminary look

■ **Imperfect competition and random technologies.** The interested reader can verify that a "unanimity theorem" can be proved whenever

$$\pi^j(d^j, \theta) = f^j(d^j) + g^j(d^j)h^j(\theta), \quad j = 0, \ldots, N, \tag{30}$$

where d^j may be a vector of decision variables.[32] Note Diamond's model with technological uncertainty fulfills this condition, as does the competitive firm facing random selling price. But if output is the only decision variable and costs remain nonstochastic, (30) also is satisfied if random demand functions take the (imperfectly competitive) form

$$p^j(q^j, \theta) = f^j(q^j) + g^j(q^j)h^j(\theta), j = 0, \ldots, N, \tag{31}$$

which includes both multiplicative random demand curves $[f^j(q^j) \equiv 0]$ and additive random demand curves $[g^j(q^j) \equiv 1]$. A value theory similar to "$MR = MC$" can be constructed, and Pareto

---

[31] If investor m holds a short position in firm k in equilibrium, his utility will be *minimized* with respect to q^k at $\hat{q}^k$. It is not clear that we have Pareto optimality in this situation: perhaps investor m could bribe the stockholders of firm k to change output in exchange for some of the riskless asset (for example), leaving all better off.

[32] Ekern and Wilson [7] show that unanimity will hold under somewhat more general conditions: when the distributions of marginal profits are spanned by the distributions of current profits. (See also note 22 above.) When profits take the form (30), conditions (8) are replaced by

$$\hat{V}^j(\hat{d}^j) = \frac{f^j(\hat{d}^j) - \left(\frac{f_k{}^j}{g_k{}^j}\right) g^j(\hat{d}^j)}{r}$$

where $f_k{}^j$ is the partial derivative of f^j with respect to the kth component of the vector d^j, evaluated at $\hat{d}^j$. This condition collapses to (19) when $f^j(d^j) = -C^j(\hat{q}^j)$ and $g^j(d^j) = \hat{q}^j$. A further paper will examine implications of equilibrium in noncompetitive environments.

optimality (for the stockholders) will follow. If $\pi^j(d^j, \theta)$ cannot be put in the form (30), then in the absence of further assumptions about mean-variance preferences or special types of securities, our "unanimity" theorem breaks down, and with it all other results in the preceding sections.[33]

☐ **A different objective function.** Earlier results assumed investors ranked alternative portfolios on the basis of expected utility of returns. While our formulation is consistent with the studies of Diamond and Stiglitz, it could be argued that final portfolio value is the appropriate argument in a world which does not self-destruct at the end of the current period. We consider now a model in which end-period wealth is the appropriate argument of the utility function.

If $\hat{V}^j$ is the equilibrium market value of firm j, $(\hat{V}^j + \pi^j)\hat{s}_i^j$ will be the assumed end-period wealth resulting from investor i holding a fraction $\hat{s}_i^j$ of firm j.[34] We can then express

$$W_i = \sum_{j=0}^{N} (\hat{V}^j + \pi^j)\hat{s}_i^j,$$

or substituting from the budget constraint (5),

$$W_i = (1 + r) \sum_{j=0}^{N} \hat{V}^j \bar{s}_i^j + \sum_{j=1}^{N} (\pi^j - r\hat{V}^j)\hat{s}_i^j.$$

The investor seeks to

$$\underset{s_i \in B_i}{\text{maximize}}\ E_i[U_i(W_i, \theta)],$$

which gives the portfolio optimizing conditions

$$E_i\left[U'_i(W_i, \theta)\frac{\partial W_i}{\partial s_i^j}\right] = E_i\{U'_i(W_i, \theta)[\pi^j - r\hat{V}^j]\} = 0, \quad j = 1, \ldots, N.$$

These conditions are identical to (8) except W_i replaces R_i. But, assuming $\hat{s}_i^j = \bar{s}_i^j$ (current portfolios are optimal),

$$\frac{\partial E_i[U_i(W_i, \theta)]}{\partial q^k} = E_i\left\{U'_i(W_i, \theta)\left[\frac{d\pi^k}{dq^k}\hat{s}_i^k + \sum_{j=1}^{N}\frac{\partial \hat{V}^j}{\partial q^k}\hat{s}_i^j\right]\right\}.$$

In general, we cannot expect a unanimity theorem. However, in the purely competitive environment, we can assume investors *perceive* (1) $\partial\hat{V}^j/\partial q^k = 0$, $j \neq k$, and (2) maximizing relative value maximizes market value.

The argument which under competitive markets showed q^k

---

[33] It can be shown that if condition (30) is not satisfied, stockholders in many situations will unanimously vote to split the firm into separate firms, each of whose profit functions will take the form (30). Thus, the condition for unanimity may not be so restrictive as it first appears.

[34] See note 17 for a discussion of when "capital gains" are included in this formulation.

maximized relative value V^k (Theorem 6) also will hold when W_i replaces R_i. Thus, q_*^k, the q^k which sets

$$E_i\left[U'_i(W_i, \theta)\,\frac{d\pi^k}{dq^k}\right] = 0,$$

also sets $\partial\hat{V}^k/\partial q^k = 0$ as perceived by the investor. Together, conditions (1) and (2) imply

$$E_i\left\{U'_i(W_i, \theta)\left[\frac{d\pi^k}{dq^k}\hat{s}_i^k + \sum_{j=1}^{N}\frac{\partial\hat{V}^j}{\partial q^k}\hat{s}_i^j\right]\right\} = 0, \quad k = 1, \ldots, N,$$

when $q^k = q_*^k$, for all investors $i = 1, \ldots, M$. Under these conditions, then, the value theory for competitive markets developed in Section 7 continues to hold, as will the Pareto optimality results of Section 8 (with W_i replacing R_i and q_*^j replacing q^j, for all i and j). The firm facing certain price p^0 will continue to choose the profit maximizing $q^0 = \hat{q}^0$.

10. A mean-variance example

Let us briefly descend to a world in which

(1) all investors rank portfolios on the basis of mean and variance of return; and
(2) all investors share identical expectations for prices, and therefore for profits.

This is, of course, the world of Sharpe, Jensen and Long, Stiglitz, and Wilson,[35] where relative market values are given by the formula[36]

$$\hat{V}^j(q) = \frac{E\pi^j - k\sum_{l=0}^{N} E(\pi^j - \bar{\pi}^j)(\pi^l - \bar{\pi}^l)}{r}, \tag{32}$$

where k is the "price of risk," $\bar{\pi}^j = E(\pi^j)$, and $\pi^j = p^j(\theta)q^j - C^j(q^j)$. For the present, let us further assume that $E(\pi^j - \bar{\pi}^j)(\pi^l - \bar{\pi}^l) = 0$ for all $l \neq j$. Then (32) becomes

$$\hat{V}^j(q) = \frac{E\pi^j - k\text{Var}\,\pi^j}{r}. \tag{33}$$

Now substitute (33) into the fundamental equilibrium relation (19):

$$\frac{C^j(\hat{q}^j) + E(p^j)\hat{q}^j - C^j(\hat{q}^j) - k\text{Var}\,[\pi^j(\hat{q}^j)]}{\hat{q}^j} = MC^j(\hat{q}^j),$$

or

$$E(p^j) - \frac{k\text{Var}\,\pi^j(\hat{q}^j)}{\hat{q}^j} = MC^j(\hat{q}^j). \tag{34}$$

Clearly (34) implies $E(p^j) > MC^j(\hat{q}^j)$, which if MC is increasing in q implies output of the competitive firm under uncertainty *is*

---

[35] In [16], [9], [17], and [20], respectively.

[36] See Stiglitz [17], p. 36.

less than output under certainty.[37] However, as the variance of profit shrinks to zero, or as the risk adjustment factor k becomes small, the firm moves closer to maximizing expected profit.

These results enable us to examine Arrow and Lind's[38] contention that risk neutrality is appropriate for assessing investment decisions. For small k, their assumption appears to be an adequate approximation. If k is not small, and/or the variance of the project under consideration is great, the planner should exhibit some degree of risk aversion when returns to the project are independent of other investments.

If a firm has a substantial negative covariance with other firms, it is possible that $\sum_{l=0}^{N} (\pi^j - \bar{\pi}^j)(\pi^l - \bar{\pi}^l) < 0$, in which case the firm would have a *larger* output under uncertainty than it would if price equalled its expected value with probability one.

11. Conclusion

Examining production in the context of a stock market, we have shown the following.

(1) Despite differences in expectations and attitudes towards risk of stockholders, outputs selected by firms will be (locally) optimal for all stockholders.
(2) Firms whose profits are not random will choose output to maximize profit.
(3) The outputs selected will be Pareto optimal for stockholders.[39]
(4) The outputs selected will not in general maximize the market value of the firm. But they will maximize market value in a "perfectly competitive" environment, in which firms in the same industry have perfectly correlated returns.

The contrasting results of Diamond, Stiglitz, and Wilson on the optimality of stock markets are readily interpretable in the context of our model. Diamond assumes the "perfectly competitive" environment, which implies that value maximization coincides with stockholders' interests. Stiglitz assumes nonperfectly correlated firms, but asserts that firms maximize market value, an assertion which Wilson shows will not coincide with stockholders' interests.

Our theory indicates the essential role of the stock market in determining production decisions under uncertainty. Equilibrium output conditions involve both production (cost) and financial (market value) variables. The equilibrium conditions provide a basis for deriving comparative static results, and seem to offer

---

[37] Note $\hat{q}^j$ does *not* coincide with the value maximizing output q_v^j, which would satisfy the condition

$$Ep^j - \frac{2k \operatorname{Var} \pi^j(q_v^j)}{q_v^j} = MC^j(q_v^j).$$

[38] In [2].

[39] See our caveat in note 31 regarding Pareto optimality with equilibrium short positions.

promise for the next task: the construction of a model which includes stochastic equilibrium in the markets for outputs, as well as equilibrium in capital asset markets.

Finally, we note that our theory generates testable conclusions which do not require estimates of subjective probabilities. In this respect, it is different from the capital asset market equilibrium models, which have required approximations of subjective probabilities of stockholders for testing. In our model, subjective probabilities help to determine relative market values, but we need only measure these market values and cost functions of firms to implement empirical tests.

References

1. Arrow, K. J. "The Role of Securities in the Optimal Allocation of Risk-Bearing." *Review of Economic Studies,* Vol. 31, No. 2 (April 1964), pp. 91–96.
2. ——— and Lind, R. C. "Uncertainty and the Evaluation of Public Investment Decisions." *The American Economic Review,* Vol. 60, No. 3 (June 1970), pp. 364–378.
3. Baron, D. "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition." *International Economic Review,* Vol. 11, No. 3 (October 1970), pp. 463–480.
4. Debreu, G. *Theory of Value.* New York: John Wiley, 1959.
5. Diamond, P. "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty." *The American Economic Review,* Vol. 57, No. 4 (September 1967), pp. 759–776.
6. Dreze, J. "A Tâtonnement Process for Investment under Uncertainty," in G. Szego and K. Shell, eds., *Mathematical Methods in Investment and Finance,* Amsterdam: North Holland, 1972.
7. Ekern, S. and Wilson, R. "On the Theory of the Firm in an Economy with Incomplete Markets," Working Paper No. 19, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1973.
8. Fama, E. "Perfect Competition and Optimal Production Decisions under Uncertainty." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 2 (Autumn 1972), pp. 509–530.
9. Jensen, M. and Long, J. "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 151–174.
10. Leland, H. "Theory of the Firm Facing Random Demand." *The American Economic Review,* Vol. 62, No. 3 (June 1972), pp. 278–291.
11. Lintner, J. "The Impact of Uncertainty on the 'Traditional' Theory of the Firm: Price Setting and Tax Shifting," in J. Markham and G. Papanek, eds., *Industrial Organization and Economic Development,* Boston: Houghton-Mifflin, 1970.
12. ———. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics,* Vol. 5, No. 1 (February 1965), pp. 13–37.
13. Long, J. "Wealth, Welfare, and the Price of Risk." *Journal of Finance,* Vol. 27, No. 2 (May 1972), pp. 419–433.
14. Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica,* Vol. 34, No. 5 (October 1966), pp. 768–783.
15. Sandmo, A. "On the Theory of the Competitive Firm under Price Uncertainty." *The American Economic Review,* Vol. 61, No. 1 (March 1971), pp. 65–73.
16. Sharpe, W. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance,* Vol. 19, No. 4 (September 1964), pp. 425–492.
17. Stiglitz, J. "On the Optimality of the Stock Market Allocation of

Investment." *Quarterly Journal of Economics*, Vol. 86, No. 1 (February 1972), pp. 25–60.

18. ———. "On the Optimality of the Stock Market Allocation of Investment," paper presented to Far Eastern Meetings of the Econometric Society, Tokyo.
19. STIGUM, B. "Competitive Equilibria under Uncertainty." *Quarterly Journal of Economics*, Vol. 83, No. 4 (November 1969), pp. 533–561.
20. WILSON, R. "Comment on J. Stiglitz, 'On the Optimality of Stock Market Allocation of Investment,'" Working Paper No. 8, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1972.

The optimality of a competitive stock market

Robert C. Merton

Professor of Finance
Sloan School of Management
Massachusetts Institute of Technology

and

Marti G. Subrahmanyam

Instructor of Finance
Sloan School of Management
Massachusetts Institute of Technology

The findings of Jensen-Long and Stiglitz on the optimality of the stock market allocation have led to a controversy over whether the sources of the nonoptimality of value maximization are noncompetitive assumptions about the capital market or are inherent externalities associated with uncertainty which do not disappear even in a competitive market. At least, for the mean-variance model with constant returns to scale technologies, we claim that the answer is the former, and that the sources of the nonoptimality are nonprice-taking behavior by firms, restrictions on the availability of technologies to firms, and restrictions on the number of firms that can enter. Using both tatonnement and nontatonnement processes, it is shown that if firms value maximize, then the Jensen-Long and Stiglitz equilibria are unstable with respect to the number of firms. If the restrictions on the availability of technologies and on the number of firms that can enter are relaxed, then the equilibrium will be a Pareto optimum in most cases, and in no case will the aggregate amount of investment be less than the Pareto optimal amount. If firms act as price takers (with or without the other restrictions), then the equilibrium is a Pareto optimum.

1. Introduction

■ In their important papers on the optimality of the stock market allocation, Jensen and Long and Stiglitz[1] use a mean-variance model of the capital market to demonstrate that the investment allocation by value-maximizing firms will not in general be Pareto optimal. While both analyses are technically correct, the interpre-

---

Robert C. Merton received the B.S. in engineering mathematics from Columbia University's School of Engineering and Applied Science (1966), the M.S. in applied mathematics from the California Institute of Technology (1967), and the Ph.D. from the Massachusetts Institute of Technology (1970). Currently, he is conducting research in capital theory under uncertainty.

Marti G. Subrahmanyam received the B.A. from the Indian Institute of Technology, Madras, in 1967, the Post Graduate Diploma in Business Administration from the Indian Institute of Management, Ahmedabad, tation of their findings has led to a controversy over whether the sources of the nonoptimality of value maximization are noncompetitive assumptions about the capital market or are inherent externalities associated with uncertainty which do not disappear even in a competitive market. If the answer is the former, then their findings are consistent with certainty theory analysis, where value or profit maximization by firms with some positive degree of monopoly power need not lead to Pareto optimal allocations. If the answer is the latter, then, indeed, the findings are more fundamental. In his comment on the Jensen-Long and Stiglitz papers, Fama[2] argues that because both analyses use noncompetitive assumptions, they do not answer the question whether in the mean-variance model, a competitive equilibrium is a Pareto optimum. Fama does perform the analysis based on what he calls the appropriate competitive assumptions and concludes that, due to inherent externalities caused by uncertainty, a competitive equilibrium will not, in general, be a Pareto optimum.

Our paper addresses the issue of whether a competitive equilibrium in the mean-variance model is a Pareto optimum or not, and its principal conclusion is that it is. The reason that our conclusion differs from the previous analyses is not technical, but rather a difference in interpretation of what the competitive market assumptions are. Because of the apparent disagreement in the literature as to what the "correct" competitive assumptions are in a model of uncertainty with incomplete markets, we consider a number of alternative assumptions about firms' beliefs and the market structure and examine the optimality of the resulting equilibria. We believe that these assumptions span the range of what most people would be willing to accept as descriptive of a "competitive" market. There are two distinct issues to be resolved: (1) if firms act as "price takers" and value-maximize (the usual competitive assumptions), is the resulting equilibrium a Pareto optimum? (2) if firms do not act as "price takers" and do value-maximize but the market structure is such that entry is free, is the resulting equilibrium a Pareto optimum?

We define a firm as a "price taker" if it perceives a horizontal supply curve for capital: i.e., it believes that the required expected return on a project (the "cost of capital") is unaffected by actions taken by the firm. As will be shown for perfectly correlated outputs, the "price taker" assumption is equivalent to the firm's taking the aggregate amount of investment in a project as "fixed" since the cost of capital depends only on the aggregate amount of investment taken by all firms in a given project. In all three of the aforementioned papers, firms do not act as "price takers" in this sense. While our definition of price-taking may be open to debate, it is difficult to imagine a definition for a competitive market

---

and the Ph.D. in finance from the Sloan School of Management, M.I.T. His current research is in capital market theory.

Aid from the National Science Foundation is gratefully acknowledged. We thank S. C. Myers and M. B. Goldman for helpful comments. We especially appreciate the many suggestions by P. A. Diamond, E. Fama, M. Jensen, and J. Long to improve an earlier draft.

[1] In [4] and [13], respectively.

[2] In [3].

where entry into that market is restricted. Since all three papers restrict entry, either explicitly in the case of Jensen and Long, and Stiglitz, or implicitly, in the case of Fama, their findings do not demonstrate that a competitive equilibrium is not a Pareto optimum.

If new firms are restricted from raising funds in the capital market on the same terms as already existing firms, then the capital market does not satisfy the pure competition assumptions. This restriction is implicitly built into a model which holds constant the number of (noncolluding) firms.[3] To restrict entry into an industry by assuming that all technologies are not freely available to all firms (both existent and potential) is also a violation of the pure competition assumptions. This restriction can be built into a model by holding constant the number of firms, or more subtly, by assuming that the correlation between returns on the "same" project taken by different firms is not perfect. Since in a portfolio context, projects and technologies, or for that matter, assets and firms, are defined by their probability distributions, it is misleading and inaccurate to refer to a project being considered by two (or more) firms as the "same" project if for the same inputs, the cash flows from the project in each state of nature are not the same for each firm (i.e., if the cash flows from the same project taken by different firms are not perfectly correlated). For example, two projects or technologies whose cash flows or outputs are identically and independently distributed should not be considered the same project for the same reason that an investor does not treat two assets whose returns happen to be identically and independently distributed as the same asset or even as members of the same "risk class."[4]

Jensen and Long are careful to acknowledge that either the assumption of holding the number of firms fixed or the assumption that the correlation between cash flows on the same project taken by different firms is not perfect may not be consistent with free entry and pure competition. Stiglitz recognizes[5] that by holding the number of firms fixed, he is restricting entry, but because he often assumes in his analysis that the distributions of returns for the "same" project across firms are independent, he is led to conclude that even if the number of firms allowed to enter is expanded, the Pareto optimal allocation is not achieved. Fama[6] shows that when the returns from projects are less than perfectly correlated across firms, there are natural externalities in their production decisions that violate the conditions of a perfectly competitive market. He then uses the linear market model which satisfies the conditions of perfect competition, to show that if the number of firms is finite and returns across firms are less than perfectly correlated, the resulting equilibrium is non-Pareto optimal. As shown later, for the technologies with imperfect correlation examined in our paper, the equilibrium solution to his equations (12)–(14) requires an infinite number of firms. For other technologies, there is some question as to the existence of a solution to his equations.

---

[3] In the usual analysis of a competitive equilibrium under certainty (See, for example, [1] or [2]), it is assumed that there is a fixed number of firms, and, therefore, entry is restricted in the sense that we use it. However, because it is also assumed that firms are price takers, the entry restriction is never binding. Hence, the equilibrium is the same as if there were no such restriction. However, when firms do not act as price takers, restricting the number of firms to any finite number can affect the equilibrium allocation and allow for non-Pareto optimal results.

[4] Risk class is defined as in Miller and Modigliani [8]: namely, two assets are in the same risk class if they are considered perfect substitutes for each other by investors.

[5] In [13].

[6] In [3].

In Section 2, we reexamine the Jensen-Long model[7] and show that if firms are price takers, then the equilibrium allocation of investment is Pareto optimal. It is also shown that if firms are not price takers but entry is free, then the equilibrium level of investment will be no smaller than the Pareto optimal amount. In Section 3, we modify the Jensen-Long model to allow for reversible investment and show that the competitive equilibrium is a Pareto optimum. In Section 4, we discuss the Jensen-Long alternative criterion of social wealth maximization. Because the strict assumption of technologies' being freely available to all firms (i.e., perfect correlation for the same project across firms) is not empirically descriptive, in Section 5, we examine the nonperfect competition case of imperfect (but positive) correlation and conclude that provided new firms can enter and raise capital on the same terms as preexisting firms, the resulting equilibrium will be a Pareto optimum.

2. Jensen-Long model

■ In this section, we use the same model as Jensen and Long: namely, all investors are risk-averse and characterize their decisions based on the mean and variance of end-of-period wealth; the market is perfect with all assets infinitely divisible; all investors have homogeneous expectations; transactions costs and taxes are zero. Under these conditions, they show[8] that the equilibrium value of the firm is

$$V_j = \frac{1}{r}[\overline{D}_j - \lambda\sigma_{jM}] \quad \text{for all } j, \tag{1}$$

where

$\overline{D}_j$ = the expected value of the total end-of-period cash flow to the shareholders of firm j,

r = $1 + i$, where i is the one-period riskless rate of interest at which every investor can borrow or lend,

$\sigma_{jM} = \sum_k \sigma_{jk}$ = covariance of the total cash flows of the firm, $\widetilde{D}_j$, with $\widetilde{D}_M$ the total cash flows from all firms,

σ_{jk} = $\text{Cov}(\widetilde{D}_j, \widetilde{D}_k)$ for all j and k,

σ^2_M = $\text{Var}(\widetilde{D}_M)$,

λ = $[\overline{D}_M - rV_M]/\sigma^2_M$ = market price per unit of risk, and

$V_M = \sum_k V_k$ = total market value of all firms.

---

[7] In [4].

[8] See [4], p. 153, (1).

We also assume that λ is a fixed number. This assumption is consistent with constant absolute risk aversion on the part of investors. Consider a new investment opportunity, or project with cash flow return per unit input, $\tilde{\rho}$, where

$$\begin{aligned} E(\tilde{\rho}) &= \bar{\rho} \\ \mathrm{Var}(\tilde{\rho}) &= \sigma_\rho^2 \\ \mathrm{Cov}(\tilde{\rho}, \tilde{D}_j) &= \sigma_{j\rho}. \end{aligned}$$

It is assumed that the distribution of $\tilde{\rho}$ is independent of which firm takes it and that there are stochastic constant returns to scale. I.e., if I_j denotes the amount of investment in the project by the jth firm, then $I_j\tilde{\rho}$ is the random variable cash flow to the jth firm as a result of taking the project and the random variable $\tilde{\rho}$ is independent of the choice of I_j. As Jensen and Long do, we assume that the cash flows from all other assets remain fixed, independent of the amount of investment made in the new opportunity. I.e., it is assumed that firms cannot change the level of investment in other projects in response to investment in the new technology. Thus, not only is investment assumed to be irreversible, but, in addition, new investment cannot be made in the "old" technologies.[9] While such an assumption is probably not very realistic, the qualitative results of the Jensen-Long model are preserved when this assumption is weakened to allow for reversible investment as is shown in Section 3.

As Jensen and Long have shown,[10] the Pareto optimal amount of investment in the new project is[11]

$$I^*_w = \max\left\{0, \frac{1}{\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda\sigma_{M\rho}\right]\right\}. \tag{2}$$

Equation (2) is derived from the condition that, at the Pareto optimal level of investment,

$$\bar{\rho} = r + \lambda\sigma_{M'\rho}, \tag{3}$$

---

[9] Thus, it is assumed that there is "putty" capital for the new technology only and "clay" for all other technologies, i.e., a type of "short-run" model. We believe this description to be consistent with that in Jensen and Long. The analysis used throughout the paper is that of comparing equilibria (comparative statics; see [1], Chapter 10). I.e., we compare the alternative equilibria resulting from the introduction of a new technology and changes in assumptions about firms' perceptions. Under the alternative interpretation of looking at a cross-section of firms for a given equilibrium, the distinction between "putty" and "clay" (long-run/short-run) is not necessary. Similarly, our notion of "new" versus "old" firms and our definition of "entry" suggests a kind of dynamics which is not necessary for examination of the conditions satisfied by firms investing in the $(N+1)$st technology for a given equilibrium. Nor is it necessary to assume that λ is a constant for this type analysis. However, since much of our discussion centers on the approaches to (stability of) the equilibria, this latter interpretation is of limited usefulness.

[10] In [4], p. 156, (5).

[11] As discussed in Section 1, a project is defined by its probability distribution, and hence, the correlation of returns on the same project across firms is perfect. Hence, the aggregate level of investment for a Pareto optimum is independent of the distribution of investment in the project across firms.

where $\tilde{D}_{M'} = \tilde{D}_M + I\tilde{\rho}$ is the total cash flow for all firms when the amount of (aggregate) investment in the new project is I and where $\sigma_{M'\rho} = \sigma_{M\rho} + I\sigma_\rho^2$ is the covariance between $\tilde{\rho}$ and the return on the new market portfolio.

Jensen and Long[12] also show that the value-maximizing level of investment in the project for the jth firm is

$$I^*_j = \max\left\{0, \frac{1}{2\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda(\sigma_{M\rho} + \sigma_{j\rho} + I'\sigma_\rho^2)\right]\right\}, \tag{4}$$

where I' is the aggregate amount of investment in the project by all firms other than the jth firm.[13]

Because the value-maximizing level of investment depends upon the other projects already taken by the firm, expression (4) seems to conflict with the notion of risk independence of projects.[14] Suppose that a project is available which will not affect the distribution of cash flows of already existing projects and the distribution of the project's cash flow is the same, no matter which firm takes the project, then the doctrine of risk independence states that (1) the market values of the project and all other assets will be the same, independent of which firm takes it; (2) the firm can determine whether to take a project or not, independent of other assets held by the firm; i.e., it can act as if it were a new firm with no other preexisting assets.

Actually, part (1) of the statement does hold for the Jensen-Long model in the sense that for a given level of aggregate investment in the project, the distribution of investment in the project across firms does not affect the market value of the project or of other assets. The reason that part (2) does not obtain can be traced to the assumption by Jensen and Long that the firm believes that its investment decision will affect the aggregate amount of investment in the project. Further, through this belief, the firm perceives that its investment decision will not only affect the cost of capital for the new project, but in addition, the cost of capital on other projects previously taken by the firm. This is in contrast to Fama,[15] whose "reaction principle" requires that a firm act as a price-taker in the sense of taking the aggregate level of investment in the project as fixed.

To see this, we use equation (1) to rewrite the formula for the value of the jth firm explicitly as a function of the amount of its own investment in the new technology, I_j, and the aggregate amount of investment by all firms in the new technology, I, as

$$V_j(I_j; I) \equiv V_j(0; I) + I_j g(I), \tag{5}$$

where

---

[12] In [4], p. 161, (18).

[13] Equation (4) is valid only for perfect correlation of returns on the project across firms which has been assumed.

[14] See S. C. Myers [9] and L. D. Schall [12] for a discussion and proofs of risk independence of projects.

[15] In [3], pp. 513–514, assumption (C6).

$$V_j(0;I) \equiv \frac{1}{r}\,[\bar{D}_j - \lambda(\sigma_{Mj} + I\sigma_{j\rho})] \tag{6a}$$

and

$$g(I) \equiv \frac{1}{r}\,[\bar{\rho} - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)]. \tag{6b}$$

$V_j(0; I)$ is the market value of all other assets owned by the jth firm, given that the aggregate investment in the project is I, and $g(I)$ is the value per unit input of investment in the new technology (the same for all firms). $I_jg(I)$ is the value of firm j's ownership in the new technology and it is the amount that it would receive if it sold its part of the project to another firm or "spins it off" as a separate firm. The total market value of the project, $\sum_j I_jg(I) = Ig(I)$, is independent of the distribution of investment across firms. The required expected return per unit investment in the project equals $\bar{\rho}/g(I)$, the same for all firms. Hence the "cost of capital" for the new project depends only on the aggregate amount of investment. Thus, for a given cost of capital, the value of the project (as a specified level of investment) is the same, independent of which firm takes it.

It will throw light on the analysis to compute the value-maximizing solution using the representation for firm value in (5). Namely, to determine the optimal level of investment to take in a new project, the firm picks $I_j = I^*_j$ so as to maximize $[V_j(I_j;I) - I_j]$, which is the market value of the firm net of the cost of inputs. For $I_j = I^*_j$ to be an interior solution, the first-order condition is

$$\frac{d[V_j(I_j;\, I) - I_j]}{dI_j} = \frac{\partial V_j}{\partial I_j} - 1 + \frac{\partial V_j}{\partial I}\,\frac{dI}{dI_j} = 0 \text{ for } I_j = I^*_j, \tag{7}$$

where

$$\frac{\partial V_j}{\partial I_j} - 1 = \frac{1}{r}\,[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)], \text{ independent of } I_j, \tag{8}$$

and

$$\frac{\partial V_j}{\partial I} = -\,\frac{\lambda}{r}\,[\sigma_{j\rho} + I_j\sigma_\rho^2]\ . \tag{9}$$

If we assume, as Jensen and Long do, that the firm takes investment by other firms in the project as fixed (i.e., $I' = I - I_j$ fixed) then, $dI/dI_j = 1$, and the value-maximizing solution I^*_j is as defined in (4). If, instead, Fama's "reaction principle" is used, then $dI/dI_j = 0$ and the optimal solution for each firm is the same. Namely, if $[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)] < 0$, the firm(s) will invest nothing; if $[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)] > 0$, each firm will be willing to invest an indefinite amount; if $[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)] = 0$, each firm is indifferent to the amount it invests. Obviously, the only equilibrium solution in this case is

$$I = \operatorname{Max}\left\{0, \frac{1}{\lambda\sigma_\rho^2}[\bar{\rho} - r - \lambda\sigma_{M\rho}]\right\}, \tag{10}$$

which is the Pareto optimum amount of investment as defined in (2). As is usual for standard competitive models, while the aggregate amount of investment in the project is determinate, the scale and hence, the number of firms investing in the project is not. Note that from (5), each firm believes that its market value expands in proportion to the amount of investment it takes. Therefore, if firms believe they cannot affect the aggregate amount of investment in the project (or equivalently, its cost of capital), then the number of firms necessary to achieve a Pareto optimal allocation is finite and can be as few as one.[16]

We now turn to the Jensen-Long case where firms act to maximize market value under the assumption that investment by all other firms, I', is fixed. Unlike the price-taker case, the resulting equilibrium allocation will depend on the number of firms that can invest in the new technology. In particular, Jensen and Long have shown that if there is a finite number of such firms, then the equilibrium allocation will not be an optimum, although it will be an optimum with an infinite number of firms.

However, as we shall argue, the assumption of a finite number of firms seems somewhat inconsistent with the notion of free entry and infinitely-divisible technologies with constant returns to scale. To allow for easier comparison with the Jensen-Long analysis and in keeping with the spirit of our analysis which is one of comparing equilibria, we will call those firms which have previous investments in the other technologies, "old" (or preexisting) firms and those firms which have no previous investments but can invest in the new technology, "new" firms. If, in equilibrium, a new firm takes a positive amount of investment, it is said to have "entered," and if it has not, then it is called a "potential" new firm.

For the moment, let us make the Jensen-Long assumption that there are a finite number $(N + K)$ of firms where the first $N(j = 1, 2, \ldots, N)$ firms are old firms and the last $K(j = N + 1, \ldots, N + K)$ are new firms. From the value-maximization assumption, it must be that in equilibrium, there can be no firm which believes that it can raise its market value by a (feasible) change in its investment.[17] Assume an equilibrium with respect to these assumptions with aggregate investment, I. Now, consider what would happen if we allowed one more (potential) new firm—the $(K + 1)$st. From equation (4), its value-maximizing investment demand would be

$$I^*_{N+K+1} = \operatorname{Max}\left\{0, \frac{1}{2\lambda\sigma_\rho^2}[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\sigma_\rho^2)]\right\}. \tag{11}$$

---

[16] Hence, as is the case for the usual competitive analysis (see note 3), if firms are price takers, then the restriction of entry by postulating a fixed number of firms has no effect on the investment allocation nor on the Pareto optimality of the allocation.

[17] This equilibrium condition holds for all the models in the paper. However, what constitutes a "feasible" change in investment varies according to the particular model being discussed.

Now, if $I^*_{N+K+1} > 0$, then the firm would enter if it could. Why does it not? Formally, because it is assumed not to exist. However, if it is also assumed that the constant-returns-to-scale technology is freely available and that both old and new firms can raise funds in the capital market on the same terms, we see no apparent reason (other than implicit restrictions to entry hidden from the model's explicit assumptions) why such a firm would not be created by a profit-seeking entrepreneur.

If, on the other hand, $I^*_{N+K+1} = 0$ (as would be the case, for example, if other firms acted as price takers), then the (potential) new firm would have no incentive to enter and the equilibrium would be unaffected by expanding the number of firms which can enter. Thus, at least with respect to stability, the assumption of a fixed number of firms does not matter if the restriction is not binding in equilibrium as is the case for the standard competitive model.

Of course, our opposition to restricting the number of firms in a competitive model is an argument over an assumption, and hence, we cannot "prove" that it is inconsistent with free entry. The method of argument is stability of the equilibrium with respect to the number of firms, and we have shown that if in an equilibrium, profit opportunities exist for additional firms if the number of firms allowed were expanded, then such an equilibrium is not stable.

Since the assumption of a fixed number of firms seems to conflict with the notion of free entry, we replace it with the more compatible assumption that there is potentially an indefinitely large number of firms. From the value-maximization assumption, a necessary condition for equilibrium is that no potential new firm will have incentive to enter. From (11), this condition implies that in equilibrium, the aggregate amount of investment must satisfy $[\bar{p} - r - \lambda(\sigma_{Mp} + I\sigma_p^2)] \leqslant 0$, or in other words, the condition for no entry by an additional new firm is that $I \geqslant I^*_w$. Thus, the market cannot be in equilibrium unless aggregate investment is at least as large as the Pareto optimal amount. This result is in sharp contrast to the previous analyses where it was deduced that the equilibrium amount of investment was less than the Pareto optimal amount.

Using two different dynamic processes of adjustment toward equilibrium, we find the distribution of investment across firms which makes the equilibrium stable. Having done so, we then answer the question under what conditions will these stable equilibria be Pareto optimal?

To determine stability, we first consider a sequential-type analysis where firms arrive in the market place in a random fashion and make an investment decision according to (4). However, once they commit themselves they cannot reverse the decision (disinvest) although they can "come back" to the market place to make more investment. On completion of this analysis, we then consider the more conventional *tâtonnement* process for reaching equilibrium. For simplicity, we assume that old firms do not invest in the project (i.e., $I_j = 0$, $j = 1, \ldots, N$) although the analysis would be virtually identical provided that $\Sigma_1^N I^*_j \leqslant I^*_w$.[18]

The analysis for the case where both old and new firms invest in the new project (again, provided $\Sigma_1^N I^*_j \leqslant I^*_w$) is performed in the Appendix.

The process for getting to an equilibrium is as follows: a new firm (chosen at random) arrives at the market place and is given the following "macro" information (in addition to the "micro" information which it already has: namely, the distribution for ρ): r, λ, and the probability distribution for the market portfolio at the "time" of the firm's arrival. "Time" is counted here as incrementing by one with each arrival of a new firm in the market place. Thus, if the current new firm making an investment decision is the kth such firm to do so, then the distribution for the market cash flow at time k is denoted by the random variable $\tilde{D}_M(k) \equiv \tilde{D}_M + \Sigma_1^{k-1} I_{N+t}\rho$ where I_{N+t} is the amount of investment in the project chosen by the tth (new) firm to arrive ($\Sigma_1^{k-1} I_{N+t}$ is the aggregate amount of investment in the project by all previously arrived firms) and $\tilde{D}_M(1) \equiv \tilde{D}_M$. By choosing the information provided in this way, the new firm need not know what investments other firms have taken which is more in the spirit of decentralized decisions by firms and an impersonal market place for raising or issuing capital. Since it is assumed that each firm acts so as to maximize market value, the optimal investment by the kth (new) firm will be

$$I^*_{N+k} = \text{Max}\left\{0, \frac{1}{2\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda\sigma_{M\rho}(k)\right]\right\}, \qquad (12)$$

where $\sigma_{M\rho}(k) \equiv \text{Cov}[\tilde{D}_M(k), \tilde{\rho}] = \sigma_{M\rho} + (\Sigma_1^{k-1} I_{N+t})\sigma_\rho^2$. Formally, (12) and (4) are the same equations. However, (12) shows that direct knowledge of other firms' investment decisions is not necessary. The process of arrivals by firms continues until no new firm would want to enter in which case the market will then be in equilibrium.

To determine the amount of investment taken by the kth new firm, we have that $\Delta_k\sigma_{M\rho}(k) \equiv \sigma_{M\rho}(k+1) - \sigma_{M\rho}(k) = I^*_{N+k}\sigma_\rho^2$ and therefore, $\Delta_k I^*_{N+k} = -\frac{1}{2} I^*_{N+k}$, or

$$\begin{aligned} I^*_{N+k+1} &= \frac{1}{2} I^*_{N+k} \\ &= \left(\frac{1}{2}\right)^k I^*_{N+1}, \end{aligned} \qquad (13)$$

where $I^*_{N+1} = \text{Max}\left\{0, \frac{1}{2\lambda\sigma_\rho^2}[\bar{\rho} - r - \lambda\sigma_{M\rho}]\right\}$. Further, the aggregate amount of investment taken by the first k firms, $I(k)$, is

$$I(k) = \text{Max}\left\{0, \frac{\left[1 - \left(\frac{1}{2}\right)^k\right]}{\lambda\sigma_\rho^2}[\bar{\rho} - r - \lambda\sigma_{M\rho}]\right\}$$

---

[18] See the Appendix.

$$= \left[1 - \left(\frac{1}{2}\right)^{k}\right] I^*_w. \tag{14}$$

By this process, it actually takes an infinite number of new firms ($k = \infty$) to reach equilibrium which also corresponds to the Pareto optimal amount of investment. We did not assume that many new firms enter, but deduced it as the result of free entry and the assumption that as long as profit opportunities exist, firms will enter. Note, however, that the convergence to equilibrium is rapid: for example, after entry by ten firms, the aggregate amount of investment taken is 99.9 percent of the Pareto optimal amount. Because of the nontatonnement nature of the approach to equilibrium, there are wealth distributional effects if trades actually take place at the "false" interim prices. However, given the assumption of a constant price of risk, we have that the equilibrium will be unaffected by these transfers of initial wealth, and it will be a Pareto optimum.

An alternative interpretation of the process could be a sequence of (unstable) Jensen-Long-type equilibria where in each subsequent equilibrium, investment taken in each previous equilibrium is treated as "clay" in the same manner as investments in other technologies are in their model. It should also be noted that the dynamic process described corresponds to the static behavior of a perfectly discriminating monopolist whose behavior, in the standard theory under certainty, will produce a Pareto optimal allocation.

We now look at the equilibrium solution achieved when the approach is by the conventional tatonnement process, but where we still allow entry by new firms as long as profit opportunities exist. We can write the equilibrium value-maximizing level of investment for the jth firm from (4)

$$\begin{aligned} I^*_j &= \text{Max}\left\{0, \frac{1}{\lambda\sigma_p^2}\left[\bar{p} - r - \lambda(\sigma_{Mp} + \sigma_{jp} + I^*\sigma_p^2)\right]\right\} \\ &= \text{Max}\,\{0, (I^*_w - I^*) - \sigma_{jp}/\sigma_p^2\}, \end{aligned} \tag{15}$$

where I^* is the equilibrium aggregate amount of investment in the project and I^*_w is the Pareto optimal amount of investment as defined in (2).

To determine the equilibrium, we start by partitioning the firms into three groups: group 1 contains all firms such that $\sigma_{jp} > 0$; group 2 contains all firms such that $\sigma_{jp} < 0$; group 3 contains all firms such that $\sigma_{jp} = 0$. Groups 1 and 2 contain only old firms while group 3 contains all new firms and in addition, any old firms whose $\sigma_{jp} = 0$. For notational simplicity and without loss of generality, we assume that there are no old firms in group 3. By renumbering firms if necessary, let the firms in group 1 be numbered $1, 2, \ldots, N_1$; the firms in group 2 be numbered $N_1 + 1, \ldots, N$; the firms in group 3 to be numbered $N + 1, \ldots, N + K$ (where K is the number of new firms which enter).

By inspection of (15), all new firms will choose the same amount of investment, $I_{new} \equiv I^*_{N+1} = I^*_{N+2} = \ldots = I^*_{N+K}$. Further, for I^* to be an equilibrium amount of aggregate investment when free entry is allowed, there must be no incentive for any firm to change its chosen level of investment. In particular, consider the $(K+1)$st (potential) new firm: from (15), it will not enter only if I^* is such that

$$\bar{\rho} - r - \lambda(\sigma_{M\rho} + I^*\sigma_\rho^2) \leqslant 0. \tag{16}$$

Hence, from (15) and (16), we have that the equilibrium amount of investment by all firms in groups 1 will be zero, i.e., $I^*_1 = I^*_2 = \ldots = I^*_{N_1} = 0$. Further, strict inequality in (16) can only obtain if $I_{new} = 0$ and no new firms enter.

Equation (16) is equivalent to the condition that $I^* \geqslant I^*_w$ with $I^* > I^*_w$ only if $I^*_{new} = 0$. As in our analysis of the nontatonnement approach to equilibrium, we have found that free entry ensures that the equilibrium amount of investment in the project will never be less than the Pareto optimal amount. A sufficient condition for $I^* = I^*_w$ is that $-\Sigma_{N_1+1}^{N}(\sigma_{j\rho}/\sigma_\rho^2) \leqslant I^*_w$. If the strict inequality holds, then new firms will enter and $KI_{new} = I^*_w - [-\Sigma_{N_1+1}^{N}(\sigma_{j\rho}/\sigma_\rho^2)]$.

Inspection of the Jensen-Long[19] formula for the equilibrium amount of investment when project returns are perfectly correlated across firms shows that if any new firms enter, an infinite number will (i.e., $K = \infty$).

Hence, as in the previous nontatonnement analysis, if at least one new firm enters, then the number of new firms entering will be infinite, and the stable equilibrium will be a Pareto optimum. Again, we emphasize that it was not assumed *a priori* that the number of firms entering was infinite, but deduced from the assumptions of value-maximizing by firms, free entry, and infinitely divisible assets. The convergence to the Pareto optimum is less rapid for the tatonnement process than in the previous case because entry by a new firm induces previously entered firms to contract their (contingent) investment, i.e., $1 \geqslant I^*/I^*_w \geqslant K/(K+1)$.

Actually, the convergence to the Pareto optimum does not even require that *all* firms are value-maximizers. Essentially, the analysis has shown that as long as the aggregate amount of investment is less than the Pareto optimal amount and as long as entry is not restricted, profit opportunities will exist and new firms will enter.

In the preceding analysis, it was shown that if, at least, one new firm enters, then the equilibrium is a Pareto optimum. However, it is possible that under our definition of free entry, the equilibrium allocation of investment can be larger than the Pareto optimum amount. Further, it is difficult to find an additional mechanism which will ensure that in all cases the equilibrium is a Pareto optimum. One mechanism which would substantially reduce the possibility for such "overinvestment" would be to assume that new firms, in addition to looking for profit opportunities for direct investment in the new technology, also consider the possibility of taking over old firms and rearranging their investment in the new technology so as to make a profit.

To see the effect of this assumption, we examine the case where the project's returns are sufficiently negatively correlated with some of the currently existing assets so that the investment taken by firms in group 2 is greater than the Pareto optimum amount. For simplicity, consider the case where group 2 contains only one firm, the Nth one, and the value maximizing $I^*_N (= I^*) > I^*_w$ which implies that $-\sigma_{N\rho} > I^*_w \sigma_\rho^2$.

---

[19] In [4], p. 163, (27).

The change in market value of the kth firm, $k = 1, 2, \ldots, N-1$, going from an "equilibrium" with aggregate investment in the project of I^*_N to one with aggregate investment equal to I^*_w (with $I^*_k = 0$ in either case) would be

$$V_k(0;I^*_w) - V_k(0;I^*_N)$$

$$= \frac{\lambda}{r}[I^*_N - I^*_w]\sigma_{k\rho} > 0, k = 1, \ldots, N-1, \qquad (17)$$

and the change in value of the Nth firm if it divested itself of the investment in the project and (through the entry of new firms) the aggregate investment in the project were I^*_w, would be

$$V_N(0;I^*_w) + I^*_N - V_N(I^*_N;I^*_N)$$

$$= \frac{\lambda}{r}[I^*_N - I^*_w][\sigma_{N\rho} + I^*_N \sigma_\rho^2] \qquad (18)$$

$$= \frac{\lambda}{2r}[I^*_N - I^*_w][\sigma_{N\rho} + I^*_w \sigma_\rho^2] < 0.$$

Now, if the total change in the market value of all existing firms were positive by going from I^*_N to I^*_w, then a profit opportunity would exist for a "new" firm to buy out all existing firms at the values associated with the I^*_N "equilibrium" and then to divest the Nth firm of its investment in the project and sell everything out at the new values associated with the I^*_w equilibrium.

By summing (17) from $k = 1$ to $(N - 1)$ and adding it to (18), we have that the condition for the sum of the change in the market values of all firms to be positive is

$$\bar{\rho} - r + \frac{\lambda}{2}\sum_{1}^{N-1} \sigma_{k\rho} > 0. \qquad (19)$$

Hence, provided that (19) is satisfied and that tender offers or mergers are allowed, the competitive equilibrium (where firms act so as to maximize their market value) will be a Pareto optimum.

3. A modified Jensen-Long model

■ In the previous section, we assumed as did Jensen and Long that investment could only take place in the new technology and that investment in "old" technologies was held fixed. Both the investment demand by firms and the Pareto optimal amount of investment were computed on the basis of this assumption. While it may be realistic to assume that investment already taken is nonreversible (except possibly through some limited depreciation), there is no apparent reason to assume that investment in previous technologies might not be expanded.

In this section, we go to the other extreme and assume that all investment is reversible, i.e. there are no fixed costs, and capital can be freely transferred from one technology to another.[20] An equivalent intertemporal assumption would be that capital depreciates completely in a single period. We choose the completely reversible assumption over the more realistic "mixed" reversible-irreversible assumption described above, not only for the obvious analytical simplifications, but also to preserve the one-period nature of the problem which would not be adequate if complete depreciation were not allowed.[21]

Because we are only interested in the Pareto optimal allocation and because we maintain the assumption of a fixed price of risk λ, we can assume without further loss of generality a single representative investor with constant absolute risk aversion.

To determine the Pareto optimal allocation, we eliminate firms from the economy and derive the optimal investment allocation to each technology by the investor. One can view the problem as that faced by an economy with a single input, wheat (seed), which can be planted in various ways so as to produce a final output, wheat (grain) at the end of the period. The probability distributions for output from each technology are given and the assumption of stochastic constant returns to scale is maintained. The investor has some given initial endowment of wheat and must choose an optimal allocation across the various technologies.

First, consider the allocation when there are N technologies with expected return per unit input, α_i, and covariances of returns per unit input v_{ij}, $i, j = 1, 2, \ldots, N$. The optimal allocation of investment will be[22]

$$I^*_j = \frac{1}{\lambda} \sum\nolimits_1^N \mu_{jk}(\alpha_k - r), j = 1, \ldots, N, \qquad (20)$$

where μ_{ij} is the ijth entry of the inverse of the $N \times N$ variance-covariance matrix of returns.

Now consider how the allocation will change if a new $(N + 1)$st technology is introduced with expected return per unit input, $\alpha_{N+1} = \bar{p}$, and with covariances of returns $v_{jN+1} = v_{jp}$, $j = 1, \ldots, N$, $(v_{N+1\ N+1} = \sigma_p^2)$. The new optimal allocation of investment will be

$$I'^*_j = \frac{1}{\lambda} \sum\nolimits_1^{N+1} \mu'_{jk}(\alpha_k - r),\ j = 1, \ldots, N + 1, \qquad (21)$$

where μ'_{ij} is the ijth entry of the inverse of the (new) $(N + 1) \times (N + 1)$ variance-covariance matrix of returns.

To investigate the changes in the optimal allocation of investment caused by the introduction of the new technology, we derive some relationships between the entries in the inverse of the "old"

---

[20] In contrast to the analysis in Section 2 (see note 9), the assumptions of this section correspond to a "long-run" model and the comparative statics is between a N-technologies "putty" equilibrium allocation and a $(N + 1)$-technologies "putty" one.

[21] See Merton [7] for a discussion of an intertemporal capital asset pricing model where supplies are taken as variable.

[22] See Merton [6], p. 1863. To avoid bothersome inequalities, it is assumed that all technologies are sufficiently productive so that $I^*_j \geqslant 0$, $j = 1, \ldots, N$.

$N \times N$ variance-covariance matrix, $\Omega^{-1} = [\mu_{ij}]$ and the inverse of the new $(N+1) \times (N+1)$ variance-covariance matrix, $\Gamma^{-1} = [\mu'_{ij}]$. Define the $N \times N$ matrix, γ; the N-vectors B and E; and the scalar D by

$$B' \equiv [\nu_{1\rho}, \ldots, \nu_{N\rho}]$$

and

$$l_{N+1} = \Gamma\Gamma^{-1} = \begin{bmatrix} \Omega & B \\ B' & \sigma_\rho^2 \end{bmatrix} \begin{bmatrix} \gamma & E \\ E' & D \end{bmatrix}$$

where l_K is the $K \times K$ identity matrix. Then, γ, E, and D must satisfy

$$\begin{aligned} \Omega\gamma + BE' &= l_N \\ B'E + \sigma_\rho^2 D &= l_1 \\ \Omega E + BD &= 0 \\ B'\gamma + \sigma_\rho^2 E' &= 0. \end{aligned} \tag{22}$$

By manipulating the conditions in (22), we have that

$$D = [\sigma_\rho^2 - B'\Omega^{-1}B]^{-1} \tag{23a}$$
$$\gamma = \Omega^{-1} + D\Omega^{-1}BB'\Omega^{-1} \tag{23b}$$
$$E = -D\Omega^{1-}B. \tag{23c}$$

Noting that $[\mu'_{1N+1}, \ldots, \mu'_{N+1\,N+1}] = [E', D]$, we have from equation (21) that the optimal amount of investment in the new technology is

$$\begin{aligned} I^*_{N+1} &= \frac{1}{\lambda}\left[D(\alpha_{N+1} - r) + \sum_1^N \mu'_{jN+1}(\alpha_j - r)\right] \\ &= \frac{D}{\lambda}\left[\alpha_{N+1} - r - \sum_{j=1}^N \left(\sum_{i=1}^N \nu_{i\rho}\mu_{ij}\right)(\alpha_j - r)\right], \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{from (23c)} \\ &= \frac{D}{\lambda}\left[\alpha_{N+1} - r - \sum_{i=1}^N \nu_{i\rho}\left(\sum_{j=1}^N \mu_{ij}(\alpha_j - r)\right)\right] \\ &= \frac{D}{\lambda}\left[\alpha_{N+1} - r - \lambda\sum_{i=1}^N \nu_{i\rho} I^*_i\right], \text{ from (20).} \end{aligned} \tag{24}$$

To compare (24) with the Pareto optimal level (2) in the previous section, we note that, as defined in Section 2, $\sigma_{M\rho} = \Sigma_1^N \nu_{i\rho} I^*_i$. Hence, rewriting (24) in terms of the notation of the previous section,

$$\begin{aligned} I'^*_{N+1} &= \frac{D}{\lambda}[\bar{\rho} - r - \lambda\sigma_{M\rho}] \qquad (25) \\ &= (D\sigma_\rho^2) I^*_w \text{ as defined in (2).} \end{aligned}$$

Hence, unless $D\sigma_\rho^2 = 1$, the Pareto optimal amount of investment in the new technology when investment in the old technologies can be changed will not be the same as when investment in the old technologies is held fixed. Noting that $B'\Omega^{-1}B \geqslant 0$ because Ω^{-1} is positive definite, we have from (23a) that

$$\begin{aligned} D\sigma_\rho^2 &= [1 - (B'\Omega^{-1}B)/\sigma_\rho^2]^{-1} \\ &= [1 - (\sum_1^N \sum_1^N \nu_{k\rho}\nu_{j\rho}\mu_{jk})/\sigma_\rho^2]^{-1} \geqslant 1, \end{aligned} \tag{26}$$

with strict equality holding if and only if $\nu_{kp} = 0$ for $k = 1, 2, \ldots, N$. Thus, the Pareto optimal amount of investment in the new technology will be larger in the unconstrained case than in the constrained one unless the new technology is uncorrelated with all other technologies.[23] (Note that while $\sigma_{Mp} = 0$ is necessary, it is not sufficient to ensure that $D\sigma_p^2 = 1$.)

Although not directly related to the question of whether a competitive equilibrium is a Pareto optimum, an important difference between the Pareto optimum solution of this section and the one found by Jensen and Long is that to compute I'^*_{N+1} requires knowledge not only of the distribution for the new technology and its relationship to the distribution of the (prenew technology) market portfolio, but also of the complete variance-covariance structure for the new technology and all the other technologies (i.e., unlike in equation (2), where one could compute the Pareto optimum with only information about the project and the aggregate market portfolio, to compute the Pareto optimum in equation (25), one would have to know the complete micro-structure for all other assets). The only statement that can be made without such information is that if $[\bar{\rho} - r - \lambda\sigma_{Mp}] > 0$, then some positive amount of investment at least as large as $[\bar{\rho} - r - \lambda\sigma_{Mp}]/\lambda\sigma_p^2$ should be taken in the new technology.

In a similar fashion, we can deduce the changes in investment in the other technologies to be

$$I'^*_j - I^*_j = -\left(\sum\nolimits_1^N \nu_{kp}\mu_{kj}\right) I'^*_{N+1} \quad j = 1, 2, \ldots, N. \quad (27)$$

As in computing I'^*_{N+1}, a knowledge of the complete micro-structure for all assets is required to compute I'^*_j. Equation (27) also demonstrates that the assumption made by Jensen and Long that the riskless asset is available in unlimited amounts is not a sufficient condition to ensure that investment in the old technologies will not change in response to the introduction of a new technology.

The important questions to be answered are: (1) if we now introduce the value-maximizing firms and allow for free entry, can the resulting equilibrium be a Pareto optimum? (2) How much information is required by the firms so that the answer to (1) is yes?

Note that because capital is freely mobile among technologies, there is no need to distinguish between new and old firms. Hence, consider the *k*th firm which only has "micro" information about the *j*th technology. Suppose it acts so as to maximize its market value using the same assumption as in the previous section: namely, it takes investment by all other firms as given. Then, it will choose its level of investment I_j^k so as to

---

[23] While, at first glance, it may not be clear why the Pareto optimal amount of investment in the new technology will always be larger in the "long-run" than in the "short-run" (independent of the covariance structure), this result is just a special case of the Le Chatelier Principle as applied to a constrained-maximum system. See Samuelson [10] for a discussion of the Le Chatelier Principle as applied in economics.

$$\begin{aligned}\text{Max }\{V^k - I_j^k\} &= \text{Max}\left\{\frac{I_j^k}{r}\left[(\alpha_j - r) - \lambda\left(\sum_1^{N+1} I_i v_{ij} + I_j^k v_j^2\right)\right]\right\} \\ &= \text{Max}\left\{\frac{I_j^k}{r}[(\alpha_j - r) - \lambda(\sigma_{Mj}(k) + I_j^k v_j^2)]\right\}\end{aligned} \tag{28}$$

where I_i is the aggregate investment by all other firms in the *i*th technology at the "time" of firm *k*'s decision and $\sigma_{Mj}(k)$ is the covariance of the return (per dollar) invested in the *j*th technology with the cash flow of all investments by other firms at the "time" of firm *k*'s decision ($\sigma_{Mj}(k)$ is defined in an analogous fashion to $\sigma_{M\rho}(k)$ in the previous section). The solution to (28) is

$$\begin{aligned}I_j^k &= \text{Max}\left[0, \frac{1}{2\lambda v_j^2}\left\{\alpha_j - r - \lambda\left(\sum_1^{N+1} I_i v_{ij}\right)\right\}\right] \quad (29) \\ &= \text{Max}\left[0, \frac{1}{2\lambda v_j^2}\left\{\alpha_j - r - \lambda\sigma_{Mj}(k)\right\}\right].\end{aligned}$$

Note that the firm uses precisely the same information available to firms in Section 2. (I.e., it did not need to know the micro characteristics of other firms or technologies.) Because we allow for free entry and exit, the market can only be in equilibrium if the I_i are such that any new firm considering entry into any technology will have no incentive to do so and similarly, no firm which has previously taken investment in any technology will have any incentive to exit. From (29), these conditions require that the equilibrium investments, I^*_i, satisfy

$$\alpha_j - r - \lambda\left(\sum_1^{N+1} I^*_i v_{ij}\right) = 0, j = 1, 2, \ldots, N+1 \tag{30}$$

or that

$$I^*_j = \frac{1}{\lambda}\sum_1^{N+1} \mu'_{ij}(\alpha_j - r), i = 1, 2, \ldots, N+1, \tag{31}$$

which are the Pareto optimal amounts of investment from equation (21).

To complete the analysis, we go to the other extreme and assume that each firm is aware of all the available technologies, athough it does not know what other firms are doing. Then, the *k*th firm will choose its investment levels, $(I_1^k, \ldots, I_{N+1}^k)$, so as to

$$\begin{aligned}\text{Max}\left\{V^k - \sum\nolimits_1^{N+1} I_j^k\right\} = \text{Max}\Bigg\{\frac{1}{r}\Bigg[&\sum\nolimits_1^{N+1} I_j^k(\alpha_j - r) \\ &- \lambda\left(\sum\nolimits_1^{N+1} I_j^k\sigma_{Mj}(k) + \sum\nolimits_1^{N+1}\sum\nolimits_1^{N+1} I_j^k I_i^k v_{ij}\right)\Bigg]\Bigg\}\end{aligned} \tag{32}$$

which can be solved to yield

$$I_i^k = \text{Max}\left[0, \frac{1}{2\lambda}\left\{\sum_1^{N+1} \mu'_{ij}(\alpha_j - r - \lambda\sigma_{Mj}(k))\right\}\right], \quad i = 1, \ldots, N+1. \tag{33}$$

As before, equilibrium would require that a new firm would have no incentive to enter which from (33) requires that

$$0 = \sum_1^{N+1} \mu'_{ij}\left[\alpha_j - r - \lambda\left(\sum_1^{N+1} I^*_k v_{kj}\right)\right], i = 1, \ldots, N+1, \tag{34}$$

where we have substituted for $\sigma_{Mj}(k)$. Noting that $\sum_1^{N+1} \mu'_{ij} v_{kj} = 0$ of $i \neq k$, and $= 1$ if $i = k$, we can rewrite (34) as

$$0 = \sum_1^{N+1} \mu'_{ij}(\alpha_j - r) - \lambda I^*_i, \quad i = 1, \ldots, N+1, \tag{35}$$

which is the Pareto optimum solution as written in (21) or (31).

Thus, even if each firm has limited information with respect to other technologies and what other firms are doing, as long as entry is free and firms value-maximize, the resulting (competitive) equilibrium is a Pareto optimum.

4. On the social wealth maximizing criterion

■ In an analysis parallel to the ones deducing the amount of investment associated with Pareto optimality and value maximization of firm value, Jensen and Long examine the social wealth maximizing criterion and derive the amount of investment that maximizes the market value of all firms. The solution can be written as

$$I^*_s = \text{Max}\left\{0, \frac{1}{2\lambda\sigma_p^2}[(\bar{p} - r) - 2\lambda\sigma_{Mp}]\right\}. \tag{36}$$

Comparing (36) with (4), the social wealth maximizing solution is the same as for a firm which is the only one with access to the new project when in addition that firm owns all the other assets in the economy (i.e., when $\sigma_{jp} = \sigma_{Mp}$). Thus, if (4) is the solution for a monopolist who has sole right to the new technology, but does not control all technologies, then one might interpret (36) as the solution for a "super" monopolist who not only has sole right to the new technology, but also owns all other technologies. If one interprets such an entity to be the government, then (4) could be termed the "private" monopoly solution and (36) termed the "public" monopoly solution.

Viewed in this light, it is not surprising that Jensen and Long found that the social wealth maximizing levels of investment differed from the Pareto optimal amount.[24]

A natural question to raise is, how "nonoptimal" is the social wealth maximizing solution? To answer this question, we consider a representative investor with a mean-variance utility function of terminal wealth, $H[E(W), \text{Var}(W)]$, with $H_1 \equiv \partial H/\partial E(W) > 0$ and $H_2 \equiv \partial H/\partial \text{Var}(W) < 0$. For analytical simplicity, suppose that there is a single (composite) risky asset with expected end-of-period cash flow, $I\alpha$ and variance of the cash flow, $I^2\sigma^2$, where I is the amount of investment in the risky asset. Then, let

---

[24] See, for example, Long [5], where he shows that wealth maximization with monopoly does not lead to a Pareto optimum.

$$\begin{aligned} \alpha_p &\equiv I\alpha/V \\ \sigma_p{}^2 &\equiv \sigma^2 I^2/V^2 \\ W_0(I) &\equiv W_0 + V - I \end{aligned} \tag{37}$$

where V is the market value of the risky asset; α_p is the expected return per dollar invested in the risky asset; $\sigma_p{}^2$ is the variance of that return; $W_0(I)$ is the initial wealth of the investor as a function of the amount of investment taken in the risky asset. For a given I, the investor's demand for the risky asset in his optimal portfolio can be written as

$$D_p = (\alpha_p - r)/\sigma_p{}^2\lambda \tag{38}$$

where $\lambda \equiv -(2H_2/H_1)$ and is taken to be constant.

By definition, $E(W) = D_p(\alpha_p - r) + rW_0(I)$ and $\text{var}(W) = D_p{}^2\sigma_p{}^2$. Hence, substituting for D_p from (38), we have that for the optimal portfolio

$$E(W) = (\alpha_p - r)^2/\lambda\sigma_p{}^2 + rW_0(I) \tag{39a}$$

and

$$\text{Var}(W) = (\alpha_p - r)^2/\lambda^2\sigma_p{}^2. \tag{39b}$$

To determine the effect on the investor's welfare of a change in I, we first note that

$$\frac{dH}{dI} = H_1\left[\frac{dE(W)}{dI} - \frac{\lambda}{2}\frac{d\text{Var}(W)}{dI}\right]. \tag{40}$$

Second, the market value of the risky asset is $V = I[\alpha - \lambda I\sigma^2]/r$ and hence,

$$\frac{dV}{dI} = \frac{1}{r}[\alpha - 2\lambda I\sigma^2]. \tag{41}$$

Combining (37), (39), and (41), we have that, along optimal portfolios,

$$\frac{dE(W)}{dI} = (\alpha - r) \tag{42a}$$

and

$$\frac{d\text{Var}(W)}{dI} = 2I\sigma^2. \tag{42b}$$

Substituting from (42) into (40) we have that

$$\frac{dH}{dI} = H_1[\alpha - r - \lambda I\sigma^2], \tag{43}$$

which is positive for $I < (\alpha - r)/\lambda\sigma^2$ reaching a maximum at the Pareto optimal level $(\alpha - r)/\lambda\sigma^2$. Therefore, investor welfare increases monotonically as investment increases until the Pareto optimal level of investment is reached.

Under what condition will investors be worse off with a "public" monopoly than with a "private" one? Inspection of (4) and (36) shows that the condition is $\sigma_{jp} < \sigma_{Mp}$. How likely is it for this condition to obtain? If I_i is the aggregate amount of investment by firm i, then $\sigma_{Mp} = \sum_1^N I_i v_{ip}$. Hence, if $\sum_{\substack{1 \\ i \neq j}}^N I_i v_{ip} > 0$, then $\sigma_{Mp} > \sigma_{jp}$. Further, if the v_{ip} are generally positive and of approximately the same size, then $\sigma_{Mp} \gg \sigma_{jp}$ since $\sum_1^N I_i \gg I_j$ for a large economy. Therefore, one might expect that the "public" monopolist will restrict investment in the new technology substantially more than would the "private" monopolist. Thus, we conclude that the social wealth maximizing rule has little value as a social welfare criterion.

5. Imperfect competition

■ In the preceding sections, it was established that given the assumptions required for perfect competition, the resulting equilibrium is a Pareto optimum when firms follow the value-maximization rule. One of the competitive assumptions is that the same techniques for production are available to all firms which implies that the returns across firms on the same project are perfectly correlated. Because of this strong assumption, one might reasonably question whether the derived optimality implications of the competitive model are descriptive of actual capital markets. To at least partially answer this question, we now examine the case of a noncompetitive stock market where it is assumed that the same technology is not available to all firms. However, we retain the assumption of free entry to the capital market (i.e., that new firms can raise capital on the same terms as old firms). Further, we assume that all firms follow the value-maximization rule, and therefore, new firms will enter as long as profit opportunities exist.

The model we use is the same one used by Jensen and Long in their section three although it is also descriptive of the models used by Stiglitz and Fama. It is assumed that there exists a set of new projects whose joint distribution for per unit cash flows is symmetric,[25] and that each firm has access to only one project. I.e., the expected cash flow per unit input for each firm's new project is $\bar{p}$, the same for all firms, and similarly, the per unit variance σ_p^2 is the same for all firms. The correlation coefficient for the cash flows between any two new projects is β, the same for all firms. (Clearly, the $\beta = 1$ case corresponds to all the new projects being the same project in which case the analysis reduces to that of the previous sections.)

While these assumptions about the joint probability distribution may seem somewhat specialized, they are descriptive of a (new) industry or product where the basic technology is freely available but where the correlation between the outputs of different firms is not perfect due to (somewhat random) differences in each firm's application of the technology or managerial talents, etc.[26]

---

[25] The random variables $X_1, \ldots, X_n$ have a symmetric joint probability distribution $P(x_1, \ldots, x_n)$ if $P(\ )$ is a symmetric function in its arguments. See Samuelson [11] for further discussion.

Jensen and Long[27] deduce that the value-maximizing investment for the jth firm can be written as

$$I^*_j = \text{Max}\left\{0, \frac{1}{2\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda(\sigma_{M\rho} + \sigma_{j\rho} + \beta\sigma_\rho^2 I')\right]\right\}$$

$$= \text{Max}\left\{0, \frac{1}{(2-\beta)\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda(\sigma_{M\rho} + \sigma_{j\rho} + \beta\sigma_\rho^2 I)\right]\right\}, \tag{44}$$

where I is the aggregate investment taken by all firms in the set of new projects and $I' = I - I^*_j$. Since we allow free access to the capital markets by new firms to raise capital to invest in their own versions of the new technology, a necessary condition for equilibrium is that I be such that no additional new, value-maximizing firms should have incentive to enter. Hence, from (44) for a new firm, the equilibrium aggregate investment must be[28]

$$I^* = \text{Max}\left\{0, \frac{1}{\beta\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda\sigma_{M\rho}\right]\right\}. \tag{45}$$

If $\sigma_{j\rho} > 0$ for $j = 1, \ldots, N$, then from (44) and (45), the equilibrium amount of investment by all old firms will be zero, and by symmetry, the equilibrium level of investment taken by each new firm will be the same and equal to I^*/K where K is the equilibrium number of new firms.[29] By substituting $I^*_{N+k} = I/K$, $k = 1, \ldots, K$, into (44) and solving, we have that for K new firms, the aggregate amount of investment will be

$$I = \text{Max}\left\{0, \frac{K}{[2+\beta(K-1)]\lambda\sigma_\rho^2}\left[\bar{\rho} - r - \lambda\sigma_{M\rho}\right]\right\}. \tag{46}$$

By comparing (46) with equilibrium condition (45), we have that the number of new firms that enter is infinite ($K = \infty$) as was the case in the $\beta = 1$ analysis of previous sections.

When the number of new firms is restricted to be no larger than K, Jensen and Long[30] derived the (constrained) Pareto optimal solution for aggregate investment to be

---

[26] Although they do not explicitly state it, we believe that Jensen and Long had this basic descriptive idea in mind when they formulated the model, particularly, since they refer to the set of projects as the "same" project although strictly for $\beta < 1$, it is not.

[27] In [4], p. 161, (18).

[28] For $[\bar{\rho} - r - \lambda\sigma_{M\rho}] > 0$, (45) is only valid for $0 < \beta \leqslant 1$. Inspection of (44) shows that for $\beta = 0$, the only "equilibrium" is $I^* = \infty$, although I^*_j is finite. While for $\beta < 0$, I^* and I^*_j are infinite. However, the $\beta \leqslant 0$ cases are empirically not relevant, particularly when the amount of investment becomes unbounded.

[29] The case where $\sigma_{j\rho} < 0$ for some old firms can be handled in a similar fashion to the analysis in Section 2.

[30] In [4], p. 164, (31).

$$I_w(K) = \text{Max}\left\{0, \frac{K+N}{[1+\beta(K+N-1)]\lambda\sigma_\rho^2}[\bar{\rho} - r - \lambda\sigma_{M\rho}]\right\} \tag{47}$$

and the optimal amount of investment for each firm to be

$$I_{wj} = \frac{I_w(K)}{(K+N)}, \quad j = 1, \ldots N, \ldots, N+K. \tag{48}$$

Therefore, the unconstrained Pareto optimum will be (for $0 < \beta \leqslant 1$)

$$\begin{aligned} \lim_{K\to\infty} I_w(K) &= \text{Max}\left\{0, \frac{1}{\beta\lambda\sigma_\rho^2}(\bar{\rho} - r - \lambda\sigma_{M\rho})\right\} \\ &= I^* \text{ from } (45). \end{aligned} \tag{49}$$

Hence, for $0 < \beta \leqslant 1$, the equilibrium amount of aggregate investment will equal the Pareto optimal amount as long as entry to the capital market is unrestricted. Further, from (48) and the symmetry of the equilibrium, the equilibrium distribution of investment across firms will be Pareto optimal.[31]

Since the model of this section is essentially the same as the ones used by Fama and Stiglitz, why did they deduce different results? Fama[32] deduces from his model that unless returns are perfectly correlated across firms, the marginal rates of substitutions for investors will not equal the marginal rate of transformation for the economy as a whole, and therefore, the "competitive" equilibrium is not a Pareto optimum. In comparing his analysis with Jensen and Long's, he goes on to state,[33] "They [Jensen and Long] interpret the result as due in part to the existence of monopolistic competition in the risky industry. But that explanation is not relevant here since the analysis in this section has assumed perfect competition. The result is that the competitive equilibrium is not Pareto optimal." Fama is mistaken in both his criticism of the Jensen-Long interpretation and his interpretation of his own model's results. Throughout our paper, it has been stressed that strict interpretation of the competitive assumptions imply that (1) returns on the same project across firms must be perfectly correlated and (2) the number of entering firms cannot be *a priori* fixed without implying restricted entry to the capital market. While Fama's model does not violate the second condition, it does require that the number of firms be finite if the non-Pareto optimal result is to obtain. However, without restrictions on entry, we have shown that the equilibrium number of firms is infinite. If he had applied this additional condition to his equation (21), he would have found that the resulting equilibrium was a Pareto optimum provided that the correlation of returns between firms is positive.

---

[31] Jensen and Long also derived this result as they let the number of firms tend to infinity. Further, they do state that as the number of firms become large, the resulting solution is a competitive equilibrium. However, they did not emphasize that value-maximization by firms plus unrestricted entry to the capital market make this the only stable equilibrium solution.

[32] In [3], p. 524, (21).

[33] Fama [3], p. 525.

Stiglitz recognizes that he is implicitly restricting entry by holding the number of firms fixed, but he concludes,[34] "Extensions of the model, allowing, for instance, for free entry of firms, reinforce the results reported here. In that case, not only will the total level of investments in the risky industries be too small, but also the number of firms will not be optimal. The economy will be operating below its mean variance frontier (allowing for a variable number of firms) even when the firms are independent."

The reasons that Stiglitz' conclusions differ from ours even when he allows for free entry is that he devotes most of his analysis to the case where returns across firms are independent (i.e. $\beta = 0$). Therefore, we reexamine our analysis for the $\beta = 0$ case. For simplicity, consider the case where there are no old firms (i.e. $N = 0$ and $\sigma_{M\rho} = 0$) and $\bar{\rho} > r$. From (44), the equilibrium amount of investment by the jth new firm is[35]

$$I^*_j = \frac{1}{2\lambda\sigma_\rho^2}(\bar{\rho} - r), \tag{50}$$

which is half the Pareto optimal amount. Note, however, that as long as $\rho > r$, new firms will continue to enter, and therefore, the number of new firms that enter will become infinite. Further, unlike in the $\beta > 0$ case, each firm will invest a finite amount, and hence, the aggregate amount of investment demanded will become infinite which implies that no equilibrium exists under the hypothesized conditions.

One source of the nonexistence is the simplifying assumption that there exist unlimited amounts of riskless borrowing at the fixed rate r. While this assumption may be reasonable when the aggregate amount of investment is finite, it is absurd when the investment demand at that rate is infinite. Therefore, one would expect that r would tend toward $\bar{\rho}$ in the limit. To see why, note that if $V_j = I_j$ (which is required in equilibrium for no further entry), then the return per dollar invested in the resulting market portfolio will have expected value $\bar{\rho}$ and variance σ_ρ^2/K where K is the number of new firms. Clearly, as $K \rightarrow \infty$, the Law of Large Numbers takes over and the market portfolio dominates the riskless asset. On the other hand, if we require that $\bar{\rho}I_j/V_j = r$ to avoid this dominance, then $V_j > I_j$, and as $K \rightarrow \infty$, the market value of assets becomes infinite. Thus, for $\beta = 0$, and fixed r, no equilibrium exists under our definition of entry, and there is no stable Stiglitz-equilibrium with respect to the number of firms.

Alternatively, we could rule out the $\beta \leqslant 0$ cases as simply unrealistic when combined with entry. While there is no natural restriction on the sign of $\sigma_{M\rho}$ and β could reasonably be less than one due to slightly different applications of the same technology, etc., it is difficult to imagine that the returns on firms entering the same industry could be distributed such that $\beta \leqslant 0$, particularly under the assumption that entrance by one firm does not affect the distribution of cash flow for another firm.

---

[34] Stiglitz [13], p. 55.

[35] See Jensen and Long [4], p. 168.

While it could also be argued that having an infinite number of firms enter is equally unrealistic, this result followed directly from the assumption of infinitely-divisible assets. Clearly, if investment were "lumpy," the number of firms entering would be finite. Further, if (as we believe) our nontatonnement approach to stability used in Section 2 is more descriptive of reality than the classical tatonnement process, then, even if a finite (and not very large) number of firms enter, the discrepancy between the aggregate amount of investment taken and the theoretical Pareto optimal amount will be quite small.

6. Conclusion

■ The cornerstones of our analysis are the assumptions that as long as firms (entrepreneurs) are value-maximizers and profit opportunities exist, then if they can enter, they will enter, and in a competitive market, entry is not restricted. We have shown that if firms act as price takers, the social welfare and value-maximization criteria are consistent. If firms do not act as price takers but the market structure is such that entry is free, then the resulting equilibrium investment allocation will be no less than the Pareto optimum amount. Further, this result still obtains for imperfect competition provided that access to the capital markets is free and that similar technologies are available to all firms entering an industry.

Our paper should not be interpreted as a criticism of the Jensen and Long, Stiglitz, and Fama articles, but rather as a clarification of whether their results apply to a competitive or "near" competitive market. By using stability analysis, we have shown that their equilibria are not stable with respect to the number of firms, and we have questioned the Stiglitz and Fama definitions of "price taker." Hence, without explicit additional assumptions of barriers to entry, we also question the assumption of a fixed number of firms by Jensen and Long and Stiglitz, and of a finite number of firms by Fama. However, their works do throw light on the important problems associated with noncompetitive markets with restricted entry.

Appendix

■ Consider the case of sequential investment decision making by firms (both old and new) that arrive in the market place in a random fashion. In general, the equilibrium distribution of investment across firms and the equilibrium total investment in the project will depend on the order in which firms arrive in the market place. However, independent of the ordering, the equilibrium total investment in the project will never be less than the Pareto optimal amount. Moreover, the equilibrium total investment in the project will equal the Pareto optimal amount for any sequence of arrivals with a partial sequence which includes all old firms with a negative covariance with the project (i.e., $\sigma_{jp} > 0$) such that that at that point in the sequence, the total investment in the project is less than or equal to the Pareto optimum.

From (4), the optimal investment for the jth firm to arrive is

$$I^*_j = \text{Max}\left\{0, \frac{1}{2\lambda\sigma_p^2}[\bar{p} - r - \lambda(\sigma_{Mp}(j) + \sigma_{jp})]\right\} \quad \text{(A1)}$$

$$= \text{Max}\left\{0, 1/2\left(I^*_w - \sum_{t=1}^{j-1} I^*_t\right) - \frac{\sigma_{jp}}{2\sigma_p^2}\right\}$$

where $\sigma_{Mp}(j) \equiv \text{Cov}(\tilde{D}_M(j), \tilde{p}) = \sigma_{Mp} + \left(\sum_{t=1}^{j-1} I^*_t\right)\sigma_p^2$.

Since an equilibrium obtains only when no firm upon arrival would invest in the project, we have from (A1) that the equilibrium amount of total investment cannot be less than the Pareto optimal amount. Otherwise, a new firm would have incentive to enter.

Inspection of (A1) shows that the later a particular firm arrives in a given sequence, the smaller the amount of investment taken in the project by that firm. It is straightforward to show that the particular sequence of firm arrivals which produces the largest amount of equilibrium total investment in the project is the one where firms arrive in the order of nondecreasing covariance of the project with its own previous projects. I.e., if (by renumbering if necessary) firms are labeled by the order in which they arrive for the maximal-investment sequence, then $\sigma_{1p} \leqslant \sigma_{2p} \leqslant \ldots \leqslant \sigma_{mp} \leqslant \ldots$. If there are no old firms with negative own covariances with the project, then in the maximal-investment sequence, the first (and only) firms to take positive investment in the project are new firms. But, this is the case examined in the text. Therefore, if there are no old firms with negative own covariances, then the equilibrium quantity of investment will be the Pareto optimal amount independent of the order of arrivals.

If there are m firms with negative own covariances and if $I(m)$ denotes the aggregate amount of investment for the first m arrivals in the maximal investment sequence, then, from (A1), we have that

$$I(m) = [1 - (1/2)^J]I^*_w - \frac{1}{2\sigma_p^2}\left[\sum_{j=1}^{J} (1/2)^{J-j}\sigma_{jp}\right], \quad \text{(A2)}$$

where J is the index of the last firm to undertake positive investment in the project ($J \leqslant m$). If $I(m) \geqq I^*_w$, then $I(m)$ is a stable equilibrium amount of investment. If $I(m) < I^*_w$, then new firms will enter, and the equilibrium amount of investment will be I^*_w. Thus, from (A2), a sufficient condition for the equilibrium amount of investment to equal the Pareto optimal amount for every sequence of arrivals is that

$$\frac{1}{2\sigma_p^2}\sum_{j=1}^{J} 2^j\sigma_{jp} \geqq -I^*_w. \quad \text{(A3)}$$

References

1. Arrow, K. J. and Hahn, F. H. *General Competitive Analysis.* San Francisco: Holden-Day, 1971.
2. Debreu, G. *Theory of Value.* New York: John Wiley, 1959.

3. Fama, E. "Perfect Competition and Optimal Production Decisions under Uncertainty." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 509–530.
4. Jensen, M. C. and Long, J. B., Jr. "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 151–174.
5. Long, J. B., Jr. "Wealth, Welfare, and the Price of Risk." *Journal of Finance*, Vol. 27, No. 2 (May 1972), pp. 419–433.
6. Merton, R. C. "An Analytical Derivation of the Efficient Portfolio Frontier." *Journal of Financial and Quantitative Analysis*, Vol. 7, No. 4 (September 1972), pp. 1851–1872.
7. ———. "An Intertemporal Capital Asset Pricing Model." Forthcoming in *Econometrica.*
8. Miller, M. and Modigliani, F. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *The American Economic Review,* Vol. 48, No. 3 (June 1958), pp. 261–297.
9. Myers, S. C. "Procedures for Capital Budgeting under Uncertainty." *Industrial Management Review*, Vol. 9, No. 3 (Spring 1968), pp. 1–19.
10. Samuelson, P. A. "An Extension of the Le Chatelier Principle." *Econometrica,* Vol. 28, No. 2 (April 1960), pp. 368–379.
11. ———. "General Proof that Diversification Pays." *Journal of Financial and Quantitative Analysis*, Vol. 2, No. 1 (March 1967), pp. 1–13.
12. Schall, L. D. "Asset Valuation, Firm Investment, and Firm Diversification." *Journal of Business,* Vol. 45, No. 1 (January 1972), pp. 11–28.
13. Stiglitz, J. E. "On the Optimality of the Stock Market Allocation of Investment." *Quarterly Journal of Economics*, Vol. 86, No. 1 (February 1972), pp. 25–60.

On the theory of the firm in an economy with incomplete markets

Steinar Ekern

Assistant Professor of Economics
Norwegian School of Economics and
Business Administration

and

Robert Wilson

Professor of Decision Sciences
Graduate School of Business
Stanford University

This article establishes conditions sufficient to ensure that a decision of a firm is judged to be desirable by any one shareholder (e.g., the firm's manager) if and only if every shareholder judges it to be desirable. One such condition is that the decision would not alter the set of distributions of returns available in the whole economy. Another is that shareholders are interested only in the mean and variance of the returns from their portfolios. The analysis allows for the possibility of incomplete markets.

1. Introduction

■ One approach to the problem of selecting decision criteria for a firm is to identify those circumstances in which a manager, if delegated the task, would have an incentive to choose the "preferred" alternatives. Supposing that the manager is also a shareholder, such a circumstance would be one in which all shareholders are necessarily in unanimous agreement as to which alternatives are preferable. Then, acting in his self-interest, the manager would of his own volition make decisions which would be endorsed unanimously by the other shareholders. This circumstance is a special case of the general problem of constructing managerial incentives analyzed by Wilson.[1]

It is well-known[2] that the problem is fully solved when there

---

Steinar Ekern received the Sivilokonom and the M.B.A. from the Norwegian School of Economics and Business Administration in 1966 and the Licenciate in managerial economics from the same institution in 1970. He was awarded the Ph.D. in decision sciences by the Stanford Graduate School of Business in 1973. His current research is in the theory of capital markets.

Robert Wilson received the A.B. (1959), the M.B.A. (1961), and the D.B.A. (1963) from Harvard University. His current research centers on multiperson decision theory.

The participation of Robert Wilson in this study was supported by National Science Foundation Grant GS-2874-A1 to the Institute of Mathematical Studies in the Social Sciences, Stanford University. The authors are indebted to Alan Kraus, Hayne Leland, Robert Litzenberger and Niels Nielsen for discussions on this topic.

[1] In [12].

[2] See, for example, Debreu [1].

is a "complete" set of markets for state-contingent claims. In this case it is in the interest of each shareholder to maximize the value of the firm. The firm's state-contingent returns are evaluated at the market prices for state-contingent claims, which for each shareholder are equal to his marginal rates of substitution for contingent income.

The problem has substance, therefore, only when there are incomplete markets for state-contingent claims, as in the formulation of Radner.[3] In the case of incomplete markets, shareholders are unable to insure against every contingency, and therefore their marginal rates of substitution for contingent income may differ. Consequently, proposed changes in the firm's state-distribution of returns may be met with a divided response among the shareholders.

The main purpose of this note is to demonstrate, nevertheless, that if the alternative decisions available to the firm would not alter the set of state-distributions of returns available in the whole economy, then the shareholders would be unanimous in their preferences. In fact, each shareholder would use the current market prices for existing securities to evaluate proposed changes in the firm's distribution of returns. This result shows that the effects of inoperative markets are limited to proposals which would change the set of state-distributions of returns available in the whole economy. Proposals which would not change this set can be evaluated in terms of the prevailing prices for existing securities. In particular, failure to obtain unanimity on a proposed project is a signal that separate incorporation of the project as a new firm would enlarge the feasible set of state-distribution of returns available to the shareholders.

An interesting example of this result occurs for the model of Diamond[4] and its generalization by Leland,[5] which is analyzed in detail in a companion paper in this volume. In this special case, the technology of the firm is such that *every* proposal would leave unchanged the set of available state-distributions of returns. Consequently, unanimity is always assured in Diamond's and Leland's models.

A further special result of some interest is a demonstration that unanimity always obtains when shareholders value only the mean and variance of their portfolios. In particular, the unanimous preference of the shareholders is not necessarily such as to maximize the market price of the firm. This phenomenon has also been demonstrated recently by Stiglitz and analyzed in successive papers by Jensen and Long and Fama.[6] Here we show that it can also be explained by an arbitrage process, following an earlier line of argument by Wilson.[7] In a companion paper in this volume, Merton and Subrahmanyam[8] demonstrate that, if firms maximize their market price, then free entry of firms will ensure productive efficiency, satisfying the unanimous preferences of shareholders. Their demonstration is essentially equivalent to ours since in the mean-variance framework every shareholder holds equal proportions of every risky firm.

---

[3] In [8].

[4] In [2].

[5] In [6].

[6] In [10], [5], and [4], respectively.

[7] In [13].

[8] In [7].

It is worth noting that security markets are only one means of risk sharing. General conditions for unanimity have been presented by Wilson.[9]

We turn now to the formulation of our model and the derivation of our results. Our model is highly simplified, and our results in this limited context are at most indicative. However, in Ekern[10] the results are demonstrated to hold in a somewhat modified form when many of the simplifying assumptions to be made below are relaxed. Also, in his note in this issue, Radner provides an alternative (and, we believe, superior) formulation in terms of the Arrow-Debreu model.[11]

2. Formulation

Our basic model is very simple. We first state its major premises [formulas (1) and (2) below] and then show how these can be derived from more fundamental assumptions.

Assume that there are several individuals indexed by $i \in I$ and several firms indexed by $j \in J$. There are just two time periods, now and then, all decisions being made now and all returns from firms occurring then, depending upon which state $k \in K$ obtains then. It will suffice to suppose that there is only one commodity, which serves as money. The return of firm j in state k is a known function $r_{jk}(x_j)$ of a decision variable x_j, which for the sake of simplicity we assume to be differentiable. Individual i selects a portfolio $s^i = (s_j^i)$ in which s_j^i is the fraction of firm j which he owns (short sales, which are allowed, correspond to negative s_j^i's). Let p_j denote the price of firm j (i.e., the price of a unit fraction).

Our first premise is that at a market equilibrium there exists for each individual i a set of weights $\omega^i = (\omega_k^i)$ for the states such that for each firm j,

$$\sum_k \omega_k^i r_{jk}(x_j) = p_j. \qquad (1)$$

For most models one derives a version of (1) as a portfolio optimality condition for each individual. The weight ω_k^i is then individual i's marginal rate of substitution between present income and future income in state k. Of course the equilibrium price p_j is determined to ensure that $\sum_i s_j^i = 1$, the equality of demand and supply, and necessarily $p_j \geqslant 0$ or the firm dissolves. It is useful to regard the vector $p = (p_j)$ of firms' prices as a function $p(x) = [p_j(x)]$ of the vector $x = (x_j)$ of firms' decisions.

Our second premise is that at a market equilibrium an individual i prefers to increase the decision variable x_j of firm j if and only if

$$s_j^i \sum_k \omega_k^i r'_{jk}(x_j) > 0. \qquad (2)$$

---

[9] In [13].
[10] In [3].
[11] In [9].

This criterion expresses the requirement that individual i prefers an increment in the decision variable x_j if it would increase his "demand price" for shares of the firm, provided he is not presently a short-seller. It agrees with the market-value criterion [i.e., $p'_j(x_j) > 0$] only under certain restrictive assumptions which are amplified below.

For those readers who find the above formulation of our basic model to be unfamiliar we provide in the next paragraphs the simplest one of its several possible derivations from more familiar premises.

Consider a model of the type employed by Irving Fisher. There is only a single commodity, which is available now either for consumption now or for investment by firms to yield returns for consumption then. Individual i is endowed with e_i units of this commodity now, and also with a fraction $\bar{s}_j^i$ of each firm j. Given the firms' investment levels $x = (x_j)$ and prices $p(x) = [p_j(x)]$, individual i chooses now consumption $c_i(x)$ and a portfolio $s^i(x) = [s_j^i(x)]$ to maximize his expected utility

$$\sum_k u_{ik}[c_i, \sum_j s_j^i r_{jk}(x_j)] f_{ik} \tag{3}$$

subject to the budget constraint

$$c_i + \sum_j s_j^i p_j(x) \leqslant e_i + \sum_j \bar{s}_j^i p_j(x). \tag{4}$$

Here, u_{ik} is individual i's utility function for consumption now and then if state k obtains, and f_{ik} is his assessed probability that state k will obtain. As noted by Radner, (4) omits consideration of firms' inputs now. Let u'_{ik} denote individual i's marginal utility of consumption then if state k obtains. Assuming sufficient mathematical regularity properties, and allowing short sales, a necessary condition for optimality is that there exists a Lagrangian multiplier $\lambda_i = \lambda_i(x) \geqslant 0$ for the budget constraint such that for each firm j,

$$\sum_k u'_{ik} f_{ik} r_{jk}(x_j) = \lambda_i p_j(x). \tag{5}$$

Ordinarily $\lambda_i > 0$ and therefore if we let $\omega_k^i = \lambda_i^{-1} u'_{ik} f_{ik}$ then (5) implies our first premise (1).

It is worth noting that if there exists a firm o whose return $r_{ok}(x_o) = r_o(x_o)$ is the same in every state, then $(\sum_k u'_{ik} f_{ik}) r_o(x_o) = \lambda_i p_o(x)$ and therefore $\sum_k \omega_k^i = p_o(x)/r_o(x_o)$; of course, by construction $\omega_k^i \geqslant 0$. One can think of ω_k^i as the price that individual i would be willing to pay now for a claim to one unit of the commodity then in state k, which in the special case of complete markets must be the same for every individual. It should be noted also that the formulation of our first premise (1) requires one to take account of the means of financing. If the firm purchases the commodity amount x_j now for investment by issuing shares, then (1) stands as it is, but if (say) it issues a riskless bond at a price $b = p_o/r_o(x_o)$ then $r_{jk}(x_j) = R_{jk}(x_j) - x_j/b$, where $R_{jk}(x_j)$ measures the gross return before bond payments in state k (a similar formulation holds for risky bonds).

Now consider a proposal to increment the investment level x_j of firm j. Let $U_i(x)$ denote individual i's maximum expected utility, given the investment vector x; i.e., $U_i(x)$ is (3) evaluated at $c_i = c_i(x)$ and $s^i = s^i(x)$. One can show directly by means of the calculus that

$$\left.\frac{\partial U_i(x)}{\partial x_j}\right|_{\bar{x}} = s_j^i(\bar{x}) \sum_k u'_{ik} f_{ik} r'_{jk}(\bar{x}_j),$$

provided that $e_i = c_i(\bar{x})$ and $\bar{s}^i = s^i(\bar{x})$; that is, provided that the system is appraised *at* an equilibrium, in which all individuals are content with their current holdings. Consequently,

$$\lambda_i(\bar{x})^{-1} \partial U_i(\bar{x}) / \partial x_j = s_j^i(\bar{x}) \sum_k \omega_k^i r'_{jk}(\bar{x}_j),$$

which implies our second premise for $x = \bar{x}$.

Thus, once the economy is in equilibrium, given an investment choice x by firms, each individual i will use (2) as the criterion by which to evaluate proposals to change the investment levels. This criterion agrees with the market-value maximization criterion only if one takes the weights ω_k^i as fixed, which is well-known to be valid only for firms which are price-takers in an economy with complete markets, in which case the weights are themselves prices common to every individual.

A variety of other formulations lead repeatedly to our basic premises (1) and (2), and therefore we adopt them as the central features of our basic model. In Section 4 we will, however, offer an alternative formulation which is more appropriate for the mean-variance framework commonly employed in the theory of capital markets.

3. Unanimity

■ The claim we made in the Introduction was that if a proposal would not alter the state-distributions of returns available in the economy, then the firm's shareholders (those individuals who own *positive* fractions) will either unanimously approve it or unanimously disapprove it. In terms of our basic model a project for firm j is simply an opportunity to increase (or decrease) the decision variable x_j. [Radner provides a more rigorous formulation of a "project" as a "feasible (local) direction of change."[12]]

Proposition: The shareholders of a firm will approve or disapprove unanimously a project which would not alter the set of state-distributions of returns available to individuals in the whole economy.

The argument runs as follows. For each firm j let $r_j(x_j) = [r_{jk}(x_j)]$, the vector of state-dependent returns; thus, $r_j(x_j)$ is the state-distribution of returns available by purchasing firm j. In the whole economy each individual can obtain any state-distribution of returns in the subspace S spanned by the set of return vectors $[r_j(x_j)]$ for all firms, subject only to his budget constraint. The project to change the decision variable x_j of firm j does not alter the feasible set of state-distributions of returns if and only if the vector $r'_j(x_j) = [r'_{jk}(x_j)]$ of marginal state-dependent returns lies in S; that is,

---

[12] In [9].

$$r'_j(x_j) = \sum_{h \in J} \alpha_h^j r_h(x_h) \tag{6}$$

for some numbers α_h^j, one for each firm $h \in J$. Consequently, employing first (6) and then (1) in the criterion (2), individual i prefers (say) to increase x_j if and only if

$$\begin{aligned} 0 < s_j^i \sum_k \omega_k^i r'_{jk}(x_j) &= s_j^i \sum_k \omega_k^i \sum_{h \in J} \alpha_h^j r_{hk}(x_h) \\ &= s_j^i \sum_{h \in J} \alpha_h^j \sum_k \omega_k^i r_{hk}(x_h) \\ &= s_j^i \sum_{h \in J} \alpha_h^j p_h. \end{aligned} \tag{7}$$

Since (7) has the same sign for every shareholder, it follows that the project will be approved or disapproved unanimously by the shareholders.

An evident special case occurs when S is in fact the whole space of state-distributions of returns, for which the usual mode of proof relies upon showing that the individuals' weights are identical and equal to the prices of state-contingent claims.

A further illustration is provided by a generalization of Diamond's[13] model, also analyzed recently by Leland and Ekern.[14] Suppose that firm j's returns have the special form $r_{jk}(x_j) = g_j(x_j) + h_j(x_j)l_{jk}$ (Diamond supposes that $g_j \equiv 0$) and that there exists a riskless firm o for which $r_{ok}(x_o) = r_o(x_o)$. Short sales are allowed. Then

$$\begin{aligned} r'_{jk}(x_j) &= g'_j(x_j) + h'_j(x_j)l_{jk} \\ &= \alpha_o^j r_o(x_o) + \alpha_j^j r_{jk}(x_j), \end{aligned}$$

where

$$\alpha_j^j = \frac{h'_j(x_j)}{h_j(x_j)} \tag{8a}$$

and

$$\alpha_o^j = \frac{[g'_j(x_j) - \alpha_j^j g_j(x_j)]}{r_o(x_o)}, \tag{8b}$$

and, therefore, the proposition implies that unanimity obtains among the shareholders.

A corollary of the criterion (7) is that, for a range of choices of the decision variable x_j by firm j which do not alter the available set of state-distributions of returns, the choice unanimously preferred by the shareholders is the one for which $\sum_{h \in J} \alpha_h^j p_h = 0$. (Of course the components α_h^j and the price p_h normally depend upon x, as in Leland's model above.) It is in this sense that the firm is required to be a "price taker" in order to achieve productive efficiency. Note that the manager of firm j actually needs only to know the components (α_h^j) and the market prices (p_h) for all firms $h \in J$; no further information about shareholders' preferences is required, provided unanimity obtains. For example, in the special case of Leland's model the optimality condition takes the simple form $\alpha_o^j p_o + \alpha_j^j p_j = 0$, using the formulas (8) above for α_o^j and α_j^j, which in his companion paper in this volume Leland interprets as "marginal cost equals average cost" in the context of his model. As Leland observes, an important ramification of this result is the consequence that firms' investment, production, and choice-of-technique decisions are affected by the market price of the firm, and indeed, by the market prices of all firms.

---

[13] In [2].

[14] In [6] and [3], respectively.

4. Mean-variance model

■ A similar analysis is applicable when the individuals are interested only in certain parameters of their portfolios. We shall show here that unanimity obtains whenever the shareholders value only the mean and variance of their portfolios. Here we suppose that individuals' probability assessments agree.

Let $R(x) = [R_j(x_j)]$ be the vector of the firms' mean returns and let $V(x) = [V_{jh}(x_j, x_h)]$ be their covariance matrix. If individual i selects the portfolio s^i, then his return will have mean $s^iR(x)$ and variance $s^iV(x)s^i$. We assume that each individual's expected utility (3) takes the special form $u_i[s^iR(x), s^iV(x)s^i]$, where we have deleted consumption now for simplicity. Let $V_j(x)$ be the jth row of $V(x)$ and assume there exists a firm o with zero variance to its return. It is then easily seen that optimality of the portfolios requires that, for each individual i and each firm j,

$$R_j(x_j) - 2\omega_i V_j(x)s^i = \left[\frac{R_o(x_o)}{p_o}\right]p_j, \tag{9}$$

where $\omega_i > 0$ is individual i's marginal rate of substitution between mean return and variance. The quantity $P = 2(\Sigma\omega_i^{-1})^{-1}$ is often called the "price of risk" because (9) implies that

$$p_j = \frac{R_j(x_j) - P\sum_{h\epsilon J} V_{jh}(x_j,x_h)}{\dfrac{R_o(x_o)}{p_o}}. \tag{10}$$

A further consequence of (9) and (10) is that each individual acquires the same fraction $s_j^i = P/2\omega_i$ of each firm j.

The portfolio optimality condition (9) corresponds to our earlier premise (1). Corresponding to the criterion (2), there is now the criterion that individual i prefers to increase the decision variable x_j of firm j if and only if

$$s_j^i[R'_j(x_j) - 2\omega_i \sum_{h\epsilon J} V'_{jh}(x_j,x_h)s_h^i] > 0, \tag{11}$$

where $V'_{jh}(x_j, x_h) = \partial V_{jh}(x_j, x_h)/\partial x_j$. N.B.: If $v_j(x_j) = V_{jj}(x_j, x_j)$, then $V'_{jj} = (1/2)v'_j(x_j)$. One obtains (11) by differentiating individual i's maximum expected utility partially with respect to x_j, as was done in Section 2 for the state model. Stiglitz[15] notes that (11) differs slightly from the criterion implied by maximization of the market value (10) when P is taken to be fixed.

We shall suppose that the submatrix of $V(x)$ corresponding to the risky firms is nonsingular. It then follows that there exists a solution α^j to the equation

---

[15] In [10].

$$V'_j(x) = \alpha^j V(x),$$

where $V'_j(x) = [V'_{jh}(x_j, x_h)]$ is the vector of marginal covariances. Hence the criterion (11) reduces to

$$0 < s_j^i[R'_j(x_j) - 2\omega_i V'_j s^i] = s_j^i[R'_j(x_j) - 2\omega_i \alpha^j V(x) s^i]$$
$$= s_j^i\left\{R'_j(x_j) - \alpha^j\left[R(x) - \left(\frac{R_o(x_o)}{p_o}\right)p\right]\right\}, \qquad (12)$$

using (9). Thus, the shareholders of each firm j will be unanimous in their response to a proposal to change the decision variable x_j.

An alternative derivation of the unanimity property is provided by substituting the share formula $s_h^i = P/2\omega_i$ into (11), which shows that (11) and (12) can be reduced to

$$s_j^i[R'_j(x_j) - P \sum_{h \in J} V'_{jh}(x_j, x_h)] > 0. \qquad (13)$$

Again it is evident that the manager of a firm needs only to know the market price of risk, P, to satisfy the unanimous preferences of the firms' shareholders. (He must, however, avoid the temptation to replace V'_{jj} by v'_j in (13), as he would do were he to seek to maximize the firm's market value.)

Another derivation of the criterion (13) is obtained by considering the market opportunities of shareholders. For the sake of simplicity, consider a risky firm j whose return is uncorrelated with all other firms. Let $\sigma_j(x_j)$ be the standard deviation of returns as a function of the decision variable x_j; i.e., $\sigma_j(x_j)^2 = v_j(x_j) = V_{jj}(x_j, x_j)$. An individual i will obtain from his shares s_o^i and s_j^i in the riskless firm and firm j a mean return of $s_o^i R_o + s_j^i R_j(x_j)$ and a standard deviation $s_j^i \sigma_j(x_j)$. Consider a proposal to change the decision variable x_j *to* $x_j + dx_j$, which would change his mean return to $s_o^i R_o + s_j^i R_j(x_j + dx_j)$ and his standard deviation to $s_j^i \sigma_j(x_j + dx_j)$. He could obtain the same mean and standard deviation with shares $s_o^i - ds_o^i$ and $s_j^i + ds_j^i$ satisfying the two equalities $(s_o^i - ds_o^i)R_o + (s_j^i + ds_j^i)R_j(x_j) = s_o^i R_o + s_j^i R_j(x_j + dx_j)$ and $(s_j^i + ds_j^i)\sigma_j(x_j) = s_j^i \sigma_j(x_j + dx_j)$. Solving these equations for ds_o^i/ds_j^i and letting dx_j go to zero, one obtains

$$\frac{ds_o^i}{ds_j^i} = \frac{R_j(x_j) - \sigma_j(x_j)R'_j(x_j)/\sigma'_j(x_j)}{R_o}. \qquad (14)$$

Now if $ds_o^i/ds_j^i > p_j/p_o$, then individual i would prefer to reorganize his portfolio rather than to increment the decision variable x_j. Hence, his criterion for preferring to increment the decision variable x_j is that

$$\frac{ds_o^i}{ds_j^i} < \frac{p_j}{p_o}. \qquad (15)$$

However, employing (10) and (14), one sees that this is just the criterion

$$R'_j(x_j) - P\sigma_j(x_j)\sigma'_j(x_j) > 0, \qquad (16)$$

which is precisely the same as the criterion (13) in this case.

Still another derivation of the criterion (13) has been given by Wilson[16] based on the requirement that the individuals share risk efficiently. For example, if each individual i has a constant Arrow-Pratt measure of risk aversion $r_i = 2\omega_i$, and the firms' returns are jointly normally distributed, then efficient risk sharing requires each firm j to act as though it has a constant measure of risk aversion $r_j = (\sum_i r_i^{-1})^{-1} = P$. Thus, each firm takes the market price of risk as its measure of risk aversion. This in turn is readily shown to imply (13).

One consequence of the preceding arguments is the conclusion that the shareholders of a firm unanimously prefer that the market price of the firm *not* be maximized, which conflicts with one's intuition. Jensen and Long, Fama, and most recently Merton and Subrahmanyam[17] have argued that various modifications of the notion of perfectly competitive markets are desirable to expiate the paradox. Here, we conclude with an example which is designed to show that the fault may lie instead with the mean-variance model itself. Consider as before a firm j whose returns are uncorrelated with all other firms. Further, suppose that $R_j(x_j) = a_j x_j$ and $\sigma_j(x_j) = b_j x_j$; viz., the firm's returns have stochastic constant-returns-to-scale. Then the criterion (16) implies that the efficient choice of the decision variable is $x_j = a_j / P b_j^2$. But then according to (10) the market price of the firm is $p_j = 0$! The scheme advocated by Merton and Subrahmanyam escapes this phenomenon only by supposing a sequence of successively smaller firms, each with a smaller market price converging to zero in the limit.

5. Summary

■ The substance of our results is the demonstration, admittedly for an overly simplified model, that even with incomplete markets a unanimous response from shareholders can be expected in many cases to firms' proposals. The simplest models, such as Diamond's, Leland's, and the mean-variance model, imply unanimity in every case. The more general model asserts that unanimity will fail to obtain only if the proposal would alter the state-distributions of returns available in the economy. In the latter case, formation of a new firm, offering new securities, provides an evident market test.

References

1. Debreu, G. *The Theory of Value.* New York: John Wiley, 1959.
2. Diamond, P. "The Role of the Stock Market in a General Equilibrium Model with Technological Uncertainty." *The American Economic Review*, Vol. 57, No. 4 (September 1967), pp. 759–776.
3. Ekern, S. *On the Theory of the Firm in Incomplete Markets.* Ph.D. dissertation, Stanford University, 1973.
4. Fama, E. "Perfect Competition and Optimal Production Decisions under Uncertainty." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 509–530.
5. Jensen, M. and Long, J. "Corporate Investment under Uncertainty

---

[16] In [11].

[17] In [5], [4], and [7], respectively.

and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 151–174.

6. Leland, H. "Production Theory and the Stock Market." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 1 (Spring 1974), pp. 125–144.
7. Merton, R. and Subrahmanyam, M. "The Optimality of a Competitive Stock Market." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 1 (Spring 1974), pp. 145–170.
8. Radner, R. "Existence of Equilibrium of Plans, Prices, and Price Expectations in a Sequence of Markets." *Econometrica*, Vol. 40, No. 2 (March 1972), pp. 289–303.
9. ———. "A Note on Unanimity of Stockholders' Preferences among Alternative Production Plans: A Reformulation of the Ekern-Wilson Model." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 1 (Spring 1974), pp. 181–184.
10. Stiglitz, J. "On the Optimality of the Stock Market Allocation of Investment." *Quarterly Journal of Economics*, Vol. 86, No. 342 (February 1972), pp. 25–60.
11. Wilson, R. "Comment" (On Stiglitz, reference [10] above). Working Paper No. 8, Institute for Mathematical Studies in the Social Sciences, Stanford University, April 1972.
12. ———. "Construction of Incentives for Decentralization under Uncertainty." *La Decision*, G. Guilbaud, ed., Paris: Centre National de la Recherche Scientifique, 1970.
13. ———. "The Theory of Syndicates." *Econometrica*, Vol. 36, No. 1 (January 1968), pp. 119–132.

A note on unanimity of stockholders' preferences among alternative production plans: a reformulation of the Ekern-Wilson model

Roy Radner

Professor of Economics and Statistics
University of California, Berkeley

This note, which was stimulated by Ekern and Wilson's study of the theory of the firm in an economy with incomplete markets, gives conditions for ex ante and ex post stockholders to be unanimous in their respective preferences among alternative production plans. This modest extension of Ekern and Wilson's analysis is facilitated by reinterpreting their model in standard Arrow-Debreu terms.

■ The situation described in the first Ekern-Wilson proposition[1] on unanimity of stockholders' preferences among production plans[2] can be formally reduced to an Arrow-Debreu model with complete markets. This reformulation provides insight into the logical structure of their model, and permits one to draw an additional conclusion. Roughly speaking, one can paraphrase the results of this note as follows. Suppose that the production possibility sets of all producers span a linear subspace of distributions of returns across states of the world, and that an equilibrium of the stock market and choice of production plans has the property that the equilibrium production plans span that subspace; then the *ex ante* stockholders in each firm are unanimous in their preferences among alternative production plans, and the *ex post* stockholders of each firm are unanimous in their preferences among (local) directions of change from the equilibrium production plan.

Suppose, as in the Ekern-Wilson paper, that there are two dates (1 and 2), one commodity at each date, I consumer-stockholders, J firms, and K alternative states of the world at the second date. A plan y_j for producer j is a $(K+1)$-dimensional vector (y_{jk}), where $-y_{j0}$ is the firm's input at date 1, and y_{jk} is the

---

Roy Radner received the Ph.B. (1945), the B.S. and M.S. in mathematics (1950 and 1951, respectively), and the Ph.D. in mathematical statistics (1956) from the University of Chicago. His current research is in the economic theory of organization and information and in the economics of higher education.

Research for this note was partially supported by the National Science Foundation, and benefited from discussions with Michael Rothschild.

[1] See [2], hereafter referred to as Ekern-Wilson. Although the results in this note are formally self-contained, the note was intended to be read in conjunction with Ekern-Wilson, where the reader will also find more motivation, and also references to previous literature on the subject.

[2] See [2], p. 175.

firm's net output at date 2 in state k ($k = 1, \ldots, K$). Similarly, if $c_i = (c_{ik})$ is a consumption plan for consumer i, then c_{i0} is interpreted as his consumption at date 1, and c_{ik} is his consumption at date 2 in state k ($k = 1, \ldots, K$).[3]

Let S be a given linear subspace of R^K, and let $S' = R \times S$, a subspace of R^{K+1}. Here S represents, as on page 176 of Ekern-Wilson, a set of "state-distributions of returns." Consider now an "Arrow-Debreu" model of production and exchange,[4] in which each consumer i has a consumption set, say C_i, and each producer j has a production set, say Y_j, and assume that *the sets C_i and Y_j are all subsets of S'*. Denote by w_i the (commodity) endowment of consumer i, and assume that

$$w_{ik} = \begin{cases} e_i, & \text{for } k = 0, \\ 0, & \text{for } k = 1, \ldots, K. \end{cases}$$

As in Ekern-Wilson, let $\bar{s}_j^i$ denote the initial share of firm j held by consumer i. Finally, let q', a vector in S', denote a "commodity price vector." An *Arrow-Debreu equilibrium* is an $(I + J + 1)$-tuple, $[(c^*_i), (y^*_j), q']$ such that:

(1) q' is in S';
(2) for every j, y^*_j maximizes $q' \cdot y_j$ on Y_j; and
(3) for every i, c^*_i is a preferred vector in the set of all c_i in C_i such that

$$q' \cdot c_i \leqslant q' \cdot w_i + \sum_j \bar{s}_j^i q' \cdot y^*_j. \tag{1}$$

(4) Supply equals demand at each date and in each state, i.e.,

$$\sum_i c^*_i = \sum_i w_i + \sum_j y^*_j. \tag{2}$$

To relate an Arrow-Debreu equilibrium to an Ekern-Wilson equilibrium, write $q' = (q_0, q)$, with q in S, write $c_i = (c_{i0}, z_i)$, with z_i in S, and write $y_j = (y_{j0}, r_j)$, with r_j in S. Consider an Arrow-Debreu equilibrium such that:

$$q_0 > 0, \tag{3}$$

$$r^*_1, \ldots, r^*_J \text{ span } S. \tag{4}$$

Because of (3), we can normalize q so that $q_0 = 1$. One can show (see the end of the note) that, because of (4), there exist numbers s^*_{ij} such that

$$z^*_i = \sum_j s^*_{ij} r^*_j, \quad \text{every } i, \tag{5}$$

$$\sum_i s^*_{ij} = 1, \quad \text{every } j. \tag{6}$$

Let $p_j = q \cdot r^*_j$. Then from the budget constraint (1), for every i,

$$c^*_{i0} + \sum_j s^*_{ij} p_j \leqslant e_i + \sum_j \bar{s}_j^i (y^*_{j0} + p_j). \tag{7}$$

This "budget constraint" is to be compared with (4) of Ekern-

---

[3] I have denoted here by c_{i0} what Ekern and Wilson denote by c_i, and my c_{ik} is their $\Sigma_j s^i_j r_{jk}(x_j)$.

[4] As in, say, Debreu [1].

Wilson.[5] Furthermore, for any $c_i = (c_{i0}, z_i)$ in C_i, there exist numbers s_{ij} such that $z_i = \Sigma_j s_{ij} r^*_j$ (since z_i is in S). Hence $(c^*_{i0}, \Sigma_j s^*_{ij} r^*_j)$ is a preferred vector in the set of all vectors $(c_{i0}, \Sigma_j s_{ij} r^*_j)$ in C_i such that

$$c_{i0} + \sum_j s_{ij} p_j \leqslant e_i + \sum_j \bar{s}_j^i (y^*_{j0} + p_j). \tag{8}$$

Compare this last budget constraint with (7). Hence $\{(c_{i0}, [s^*_{ij}]), (p_j)\}$ is an Ekern-Wilson equilibrium, given the plans $y^*_j = (y^*_{j0}, r^*_j)$.

In the context of the Arrow-Debreu formulation of the model, given the price vector $q' = (1, q)$, all of the *ex ante* shareholders of firm j (i.e., all consumers i for which $\bar{s}_j^i > 0$) would agree that the firm should maximize the value $y_{j0} + q \cdot r_j$. Hence *all of the ex ante stockholders of firm j are unanimous in their preferences among production plans of the firm.*

With regard to the *ex post* stockholders, an argument similar to the one given in Ekern-Wilson can be used to show that *they are unanimous in their preferences among feasible (local) directions of change from the equilibrium production plan,* provided that the consumers' preferences are sufficiently regular (smoothness, convexity, etc.). To be more precise, for any firm h, let $f_h = (g_{h0}, g_h)$ be a vector in R^{K+1} such that, for all real numbers t in some interval $[0, t^*]$,

(1) $y^*_h + tf_h$ is in Y_h;

(2) for every i, $\gamma_i(t) \equiv (c^*_{i0}, z^*_i + ts^*_{ih} g_h)$ is in C_i.

The vector f_h is a feasible direction of change from the production plan y^*_h, and g_h is the corresponding direction of change from the return vector r_h^*. If $q \cdot g_h > 0$, then there is some sufficiently small t such that each *ex post* stockholder i for which $s^*_{ih} > 0$ prefers the consumption vector $\gamma_i(t)$ to the consumption vector c^*_i. An analogous statement (in reverse) follows if $q \cdot g_h < 0$. The logic behind this proposition is, of course, the same argument as that used to justify the standard real income test (in the small) for a single consumer.

Note that each *ex ante* stockholder with positive shares wants the firm to maximize $y_{j0} + p_j$, whereas each *ex post* stockholder with positive shares wants the firm to maximize p_j, which could be interpreted as the stock market value of the firm. Hence these two groups agree with each other after the net input of the firm at the first date has been fixed.[6]

It remains to justify the representation (5)-(6). First consider the special case in which $r^*_1, \ldots, r^*_J$ are linearly independent. Since they span S, there is a unique representation of each z^*_i in the form (5). Adding (5) over all consumers gives

$$\sum_i z^*_i = \sum_j \left(\sum_i s^*_{ij}\right) r^*_j. \tag{9}$$

But by (2), the condition that supply equal demand,

$$\sum_i z^*_i = \sum_j r^*_j. \tag{10}$$

---

[5] To make the comparison exact, we must make the further assumption that, for all j, $y_{j0} = 0$ for every y_j in Y_j. This assumption seems to be implied by (4) of Ekern and Wilson.

[6] See note 5, *supra.*

If one compares (9) and (10), and recalls that the r^*_j are linearly independent, then condition (6) follows. In the case in which the r^*_j are not linearly independent, one chooses a linearly independent subset that spans S, and then proceeds with an obvious variation on the above argument.

References

1. Debreu, G. *Theory of Value.* New York: John Wiley, 1959.
2. Ekern, S. and Wilson, R. "On the Theory of the Firm in an Economy with Incomplete Markets." *The Bell Journal of Economics and Management Science,* Vol. 5, No. 1 (Spring 1974), pp. 171–180.

SHORT ARTICLES, COMMENTS AND REVIEWS

187 Space-time pricing for public utilities
J. CRAVEN

196 Corporate diversification and market structure
C. H. BERRY

205 The CAB's domestic passenger fare investigation
G. W. DOUGLAS and J. C. MILLER, III

223 Peak-load pricing: some further remarks
O. E. WILLIAMSON

229 Depreciation in a simple regulatory model: comment
P. B. LINHART

233 Depreciation in a simple regulatory model: reply
B. L. JAFFEE

235 Cable, cities, and copyrights
B. M. MITCHELL in association with R. H. SMILEY

264 A reexamination of the prophecy of doom for cable television
R. W. CRANDALL and L. L. FRAY

290 Cowling's *Market Structure and Corporate Behavior*
L. WEISS

294 Posner's *Economic Analysis of Law*
P. A. DIAMOND

301 Noll, Peck, and McGowan's *Economic Aspects of Television Regulation*
S. M. BESEN and B. M. MITCHELL

320 Approving nuclear power plants: scientific decisionmaking or administrative charade?
P. L. JOSKOW

Space-time pricing for public utilities

John Craven

Lecturer in Economics
Eliot College
University of Kent at Canterbury

This paper presents a model of pricing for public utilities where demand varies with space, rather than with time. After different policies regarding zoning are accounted for, a model integrating the spatial and temporal models is outlined to permit an examination of the interaction of peak demands and zoning. Special assumptions on costs, capacity constraints, and the nature of the objective function are made.

1. Introduction

■ Much of the theoretical literature on problems of pricing by public utilities centers on the model where demand varies with time, and hence upon the peak-load problem.[1] It is clear, however, that a public utility may supply a product for which the demand varies other than with price or time. In this paper we shall outline a model where demand varies with "space"[2]—and, stylized as it may be, see how it differs from the time dependent problem of our earlier work.[3] The demand for a product can vary with "space" in a number of ways: for example, electricity produced at a central plant will be demanded at various distances in various amounts; telephone calls made from a city center have various destinations; rail travellers demand journeys of various lengths. In some cases it may be illegal or impractical for the pricing policy of the utility to account for spatial differences in demand and production cost; in others it is standard practice to divide the area supplied into zones and make different charges in different zones (or for journeys to different zones). After considering this zoning problem, we shall outline a model which integrates the spatial and temporal models so that the interactions of peak demands and zoning can be examined.

We shall obviously need to make many special assumptions on costs, capacity constraints and the nature of the objective function for the utility. In a short paper these can be discussed only briefly—there is clearly much remaining to be considered in this field. We can only hope that models such as ours, designed to illustrate a specific feature, can have implications in more general and more realistic models of the behavior of public utilities.

I am indebted to referees of an earlier version of this paper for their comments.

[1] See, for example, Williamson [4] and Craven [1].

[2] "Space" need not refer to a Euclidean measure of distance—our theory can cover any parameter with similar effects on demand.

[3] In [1].

2. Demand

■ Our public utility faces a demand for its product at distance r given by $q[p(r), r]$, where $p(r)$ is the price charged for the product at distance r, and such that $q \geqslant 0$ for all nonnegative prices $p(r)$ and for all distances r between 0 and some outer limit R; q is continuous everywhere and $\frac{\partial q}{\partial p}$, $\frac{\partial q}{\partial r}$ exist at all points except perhaps where q becomes zero. When $q > 0$, $\frac{\partial q}{\partial p} < 0$.

Such a formulation is immediately restrictive, for it implies that demand at distance r_1 is independent of the price charged at distance r_2: no consumer can relocate to take advantage of a lower price elsewhere. This assumption, at least in the short period is perhaps reasonable, and certainly more so than the analogous one made in the peak-load problem, where no intertemporal shifting of demand is allowed.[4]

For future use, the inverse of the demand function q at a given point r is $p(q, r)$.

3. Surplus

■ The consumers' surplus generated at point r is given by the usual expression of the area under the demand curve less the total consumer expenditure at r. That is

$$CS(r) = \int_{z=0}^{q[p(r),r]} p(z, r)\,dz - p(r)q[p(r), r].$$

Total consumers' surplus generated over the range $0 \leqslant r \leqslant R$ is

$$CS = \int_{r=0}^{R} CS(r)\,dr.$$

The utility itself achieves a surplus

$$\begin{aligned} PS &= \text{Revenue} - \text{Costs} \\ &= \int_{r=0}^{R} p(r)q[p(r), r]\,dr - C, \end{aligned}$$

where C is a function of the supplies $q[p(r), r]$ and of any overhead elements.

Total surplus is then

$$\begin{aligned} S &= CS + PS \\ &= \int_{r=0}^{R} dr \int_{z=0}^{q[p(r),r]} p(z, r)\,dz - C. \end{aligned}$$

We shall assume that our utility aims to maximize S, just as in Williamson and Craven.[5] As a measure of social welfare this has well-known flaws including the implicit constancy of the marginal utility of income and indifference to the distribution of welfare between consumers. It is possible to work from individual utility functions,[6] and to use a social welfare function that avoids these problems. Apart from problems of measurement inherent in this step into utilitarianism, one immediately becomes tempted to dispose of distributional problems through "lump sum taxes,"[7] which are much quoted by economists, but hardly known to planners. Only if the utility is placed in a full general equilibrium context, and the problems of "second best" included, can we adequately dispose of these objections.

---

[4] Cf. [4] and [1].

[5] In [4] and [1], respectively.

[6] Cf. Mohring [3].

So, we have to choose a system of prices $p(r)$ to maximize S subject to certain constraints (see below). We shall assume that only a certain number (I) of discrete values can be given to $p(r)$, each one charged in a certain zone U_i.[8] So, for $r \in U_i$ $(i = 1, \ldots, I)$, $p(r) = p_i$. We have to choose the p_i and the U_i to maximize S, which can now be written[9]

$$S = \sum_{i=1}^{I} \int_{U_i} dr \int_{z=0}^{q(p_i,r)} p(z, r)\,dz - C.$$

4. Constraints

■ Once again following the pattern of the temporal model, we shall assume that demand is limited by productive capacity. We shall make the heroic assumption that it is total supply that is thus limited, rather than supply to any particular distance. For an electric company this requires that it is the generating equipment that imposes the constraint, not the power lines to the consumers; for a telephone company it is central equipment in the telephone exchange that becomes overloaded before the lines themselves. Our constraint can be written

$$\sum_{i=1}^{I} \int_{U_i} q(p_i, r)\,dr \leqslant Q.$$

While this is likely to be the best form in certain cases, the most obvious alternative,

$$q(p_i, r) \leqslant Q(r) \quad \text{for} \quad 0 \leqslant r \leqslant R,$$

is a close parallel to the constraint of the temporal problem in our earlier work,[10] and a simple extension of that analysis should suffice to discover its properties.

We are assuming that demand is to be constrained by a suitable choice of prices and zones, rather than by leaving some consumers unsatisfied (by power cuts). The latter method of containing excess demand has considerable implications for the distribution of welfare with which our simple surplus maximizing model can hardly expect to cope. It would also lead to the consideration of demand functions with essentially random components, and further into the choice of "safety margins" of capacity, which is beyond our scope here.

5. Costs

■ In a short-period problem, where the capacity of the utility and the number of different prices to be charged are invariable, the costs C will vary only with actual supply. In general, the utility will have a complex pattern of joint supply, and so variable costs (VC) will be some general functional

---

[7] *Ibid.*

[8] As with the time periods in Craven [1], each zone may consist of several disjoint parts.

[9] We write $\int_{U_i}$ for $\int_{r \in U_i}$ to simplify notation.

[10] In [1].

$$VC = \varphi[q(p, r), r: \quad 0 \leqslant r \leqslant R],$$

so that the cost of supply of a unit of product to r will depend on the supply to other distances. It is possible that such interdependence does not exist, so that φ is separable, and thus

$$VC = \sum_{i=1}^{I} \int_{U_i} \theta[q(p_i, r), r] dr.$$

For long-period problems, we must include in the cost function the variation in cost with Q and with the number of zones I. If all costs are separable, which is the assumption that we shall make, the expression for total cost becomes

$$C = \sum_{i=1}^{I} \int_{U_i} \theta(q, r) dr + \beta(Q) + \alpha(I).$$

Substituting this into the expression for S, we can determine the necessary conditions for an optimum, and some of their implications.

6. Necessary conditions for an optimum

■ (1) Williamson shows that no price is less than marginal cost,[11] and a similar result appears in our model from the condition that $\dfrac{\partial S}{\partial p_i}$ must be nonpositive.[12] For, if it is positive, some price p_i can be increased, surplus will then increase, and the capacity constraint cannot possibly be violated by the reduction in demand in zone U_i. Now

$$\frac{\partial S}{\partial p_i} = \int_{U_i} \left(p_i - \frac{\partial \theta}{\partial q} \right) \frac{\partial q}{\partial p} dr \leqslant 0$$

and since $\dfrac{\partial q}{\partial p} < 0$ everywhere, this yields

$$p_i \geqslant \int_{U_i} \frac{\partial \theta}{\partial q} \frac{\partial q}{\partial p} dr \Bigg/ \int_{U_i} \frac{\partial q}{\partial p} dr. \qquad (1)$$

The optimal price is never less than a weighted average of the marginal costs $\dfrac{\partial \theta}{\partial q}$ where the weight at each point r is $\dfrac{\partial q}{\partial p} \Big/ \int_{U_i} \dfrac{\partial q}{\partial p} dr$.

(2) In a short period problem with fixed Q and I, the expression (1) will hold with equality if the capacity constraint is not violated by the p_i so defined. If, with equality in (1), the capacity

---

[11] In [4].

[12] In deriving first-order conditions, we can, of course, assume that all other variables have their optimal values, thus separating the problem into several distinct maximizations.

constraint is violated, it is useful to find $\frac{dS}{dQ}$, which is the total increase in surplus brought about by an increase in capacity. If Q rises to $Q + dQ$, some price can fall to $p_i - dp_i$, and given that (1) holds with strict inequality, surplus will rise by $-\frac{\partial S}{\partial p_i} dp_i$, where dp_i is given by

$$-dp_i \int_{U_i} \frac{\partial q}{\partial p} dr = dQ.$$

In zone i then, there is a trade off λ_i between surplus generated and capacity demanded which is given by

$$\lambda_i = \frac{\partial S}{\partial p_i} \Bigg/ \int_{U_i} \frac{\partial q}{\partial p} dr,$$

which, when simplified, is the difference between the two sides of (1).

The trade-off obtained by reducing p_i must be the same as that obtained by reducing any other price p_j. For, if $\lambda_i > \lambda_j$, we can, without increasing the capacity Q, increase p_j, release capacity dQ and reduce surplus by $\lambda_j dQ$. The released capacity dQ can be used to reduce p_i and thus increase surplus by $\lambda_i dQ > \lambda_j dQ$. The initial position cannot be optimal, and hence the necessary condition that holds when the capacity constraint does hold with equality is

$$\lambda_i = \lambda \quad \text{for} \quad i = 1, \ldots, I. \tag{2}$$

From this, we can see that, if the optimal price in some zone exceeds the weighted average of marginal costs found above, then in every zone, the optimal price is greater than this average by a similar amount λ.

(3) In a long period problem, an increase in Q will clearly yield extra surplus λdQ, and from the cost function, will cost $\frac{d}{dQ}\beta(Q)$, and hence the optimal value of Q is achieved when

$$\frac{d\beta}{dQ} = \lambda. \tag{3}$$

Each of the above results could have been found by applying the Kuhn-Tucker theorem,[13] but the next part, where we find the optimal conditions for the zones U_i, yields more easily to the piecemeal approach.

(4) If a small interval $(r, r + dr)$ is moved from its optimal zone U_i into zone U_j, where $p_i > p_j$, there will be a gain of surplus dS_1 and an increase of capacity demanded of dQ_1. This capacity can be released by increasing some price, p_k, thus losing surplus λdQ_1. So, if the optimal zone for $(r, r + dr)$ is U_i, dS_1/dQ_1 must be less than or equal to λ, else the increase in p_k can give rise to an increase in surplus. If $p_j > p_i$, dQ_1 capacity is released and dS_1

---

[13] See [2].

surplus is lost. In this case, it must not be possible to use this capacity to generate more than dS_1 surplus elsewhere. So, $\dfrac{dS_1}{dQ_1}$ must be greater than or equal to λ. Formally, if $(r, r + dr)$ is in zone U_i when everything is optimal, then,

$$0 \leqslant \frac{dS_1}{dQ_1} \leqslant \lambda \text{ for all } U_j \text{ with } p_j < p_i \text{[14]}$$

and

$$\frac{dS_1}{dQ_1} \geqslant \lambda \text{ for all } U_j \text{ with } p_j > p_i, \tag{4}$$

where

$$dS_1 = \left| \left\{ \int_{z=q(p_i,r)}^{q(p_j,r)} p(z,r)\,dz - [\theta(q(p_j,r),r) - \theta(q(p_i,r),r)] \right\} \right| dr$$

and

$$dQ_1 = |[q(p_j,r) - q(p_i,r)]|\; dr.$$

(5) We can define a boundary of U_i to be a point $r \in U_i$ such that every neighborhood of r contains a point in some other zone U_j. Remembering footnote 8, there may be several boundary points in a zone, but reasonable behavior of the demand and cost functions should rule out the possibility of a large number of such boundary points. If r is a boundary point between U_i and U_j, then the ratio dS_1/dQ_1 in (4), calculated for these two zones is equal to λ. For, if $p_i > p_j$ and $\dfrac{dS_1}{dQ_1} < \lambda$, then there is a point of U_j, very close to r, which does not obey the second condition of (4) with the roles of i and j reversed. So, for any r that is a boundary point between U_i and U_j,

$$\frac{dS_1}{dQ_1} = \lambda. \tag{5}$$

At such a point, of course, the utility is indifferent whether r is placed in U_i or U_j.

(6) In the long period, it may be possible to alter the number of zones I by changing metering equipment and the like. Since I can only be integer valued, we cannot use the methods of the infinitesimal calculus to discover necessary conditions for I to be optimal. However, analogous to the condition in our earlier paper[15] on the optimal number of time periods, we can develop a sufficient condition to tell us whether the present value of I should be increased.

Suppose that we have the optimal solution with a given number I of zones. We can then take some U_i and split it optimally into two parts, each with its own price. The rules for this are exactly (1) through (5) above, with the interval $[0, R]$ replaced by U_i, which must be split into two "subzones." One of the prices chosen to replace p_i will in general be greater than p_i, and the other will be less, in such ratio that the overall capacity constraint is not violated when the new zones and prices are combined with the U_j and p_j for $j \neq i$. There will be a change in surplus from this operation, involving a change in consumers' surplus, variable cost and the administrative costs of the new, and presumably more complex price structure. This last change will be given by

---

[14] $\dfrac{dS_1}{dQ_1}$ must be nonnegative, else we could simultaneously increase surplus and reduce capacity demanded.

[15] In [1].

$$\alpha(I+1) - \alpha(I).$$

If the overall change is found to be positive, I should be increased to at least $I+1$ (if n zones can be split advantageously, yielding extra surplus after taking account of $\alpha(I+n) - \alpha(I)$, the number of zones should be increased by at least n). For then, when the new overall optimum is calculated, the total surplus must be at least as great as in the "experiment" of splitting U_i, for such an experiment is a feasible proposition for the new problem. On the other hand, if no such experiment works, we have no information on whether to increase I, for the optimum with $I+1$ zones may still be greater than that with I zones.

7. Profit and loss

■ Our formulation allows no limitation on the profit or loss that will be made when prices, capacity and zones are chosen optimally. From the necessary conditions for an optimum, it is easily shown that profit (π) is given by the expression

$$\pi = \sum_{i=1}^{I} \int_{U_i} [q(p_i, r) \left(\lambda + \int \frac{\partial \theta}{\partial q} \frac{\partial q}{\partial p} dr \Big/ \int \frac{\partial q}{\partial p} dr \right) - \theta(q, r)] dr - \beta(Q) - \alpha(I).$$

The first term, $\Sigma \int q(p_i, r) \lambda dr$ is equal to $Q\lambda$ if the capacity constraint holds with equality, which will be so in the long run, assuming only that extra capacity is never free. The profit expression then falls naturally into two parts: the first given by

$$\lambda Q - \beta(Q) - \alpha(I); \tag{6}$$

and the second by

$$\sum_{i=1}^{I} \int_{U_i} \left[q \cdot \int \frac{\partial \theta}{\partial q} \frac{\partial q}{\partial p} dr \Big/ \int \frac{\partial q}{\partial p} dr - \theta \right] dr. \tag{7}$$

If the function $\beta(Q)$ exhibits constant returns, so that $\beta(Q) = bQ$ for some positive constant b, then condition (3) gives $\lambda = b$, and expression (6) reduces to $-\alpha(I)$, the administrative cost of the system with I zones. The premium λ of prices above the average marginal cost in a zone is just sufficient to offset capacity costs. The loss in (6) will be greater than $\alpha(I)$ if $\beta(Q)$ exhibits increasing returns, and less with decreasing returns.

The expression (7) may be positive or negative depending on the form of θ. It will be zero if variable cost is independent both of distance and quantity, so that $\theta = g \cdot q$ with g a positive constant (as might be the case with operator-connected telephone calls or the central generation of electric power). In this case, marginal cost pricing just covers total variable cost, and the premium λ left to deal with capacity and administrative costs.

A result of similar nature was found by Williamson[16] (whose peak-load model ignored administrative costs). In his model, time periods are given, and demand does not vary with time within them. With constant returns, the utility exactly breaks even. The model of Craven[17] has demand varying continuously with time, and constrained to be less than capacity at each instant. In that model, a similar result does not hold, except when the number of time periods increases without limit so that only those instants when demand equals capacity are charged at more than marginal cost.

We can reinstate Williamson's result here because of the difference between our capacity constraint and that of our earlier work.[18] All points of space contribute "pressure" upon capacity, and prices in each zone reflect this through the uniform excess (λ) of price over average marginal cost in the zone, which in turn contributes to cover the capacity costs.

8. Mixed space-time models

FIGURE 1

GENERAL PARTITION OF [0 R] X [0 1]

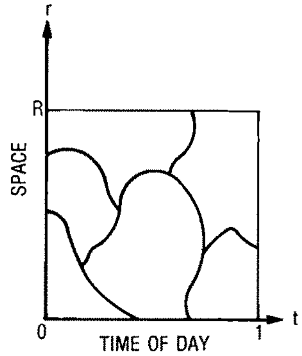
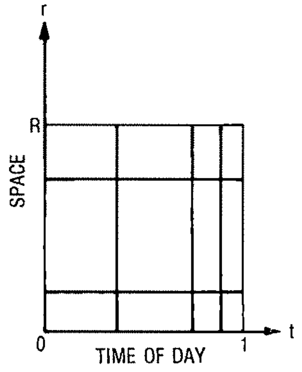


FIGURE 2

RECTANGULAR PARTITION OF [0 R] X [0 1]



■ If it is efficient for a utility to price according both to time and to distance, we shall need a combined theory, for the two problems will generally interact to become one of choosing optimal "space-time blocks," or subsets of the Cartesian product of $[0, R]$ and the "day." In general the partition could be made in the manner of Figure 1, where each area represents a subset of the space-time plane in which the price is constant. It is unlikely that the administration of such a scheme would be feasible, as there are no simple rules to tell us when to switch from one block to another, so that the problem is more likely to involve a division in the manner of Figure 2.

To combine the spatial and temporal models, we could postulate a demand function, $q(p, r, t)$ continuous and nonnegative in the relevant domain, and with $\frac{\partial q}{\partial p}$ negative. This formulation assumes that demand at a certain place and time is independent of the price charged at any other space-time coordinate. Thus no spatial or temporal shift in demand to take advantage of lower prices is possible. The inverse of q at a point (r, t) is once again the function p, and so the expression for surplus is

$$\int_{r=0}^{R}\int_{t=0}^{1}\int_{z=0}^{q(p,r,t)} p(z, r, t)\,dz\,dt\,dr - \text{Costs}.$$

If we are to choose I zones U_i and K periods T_k so that within the block V_{ik} $(= U_i \cap T_k)$ the price charged is held constant at p_{ik}, the expression for surplus can be written

$$\sum_{i}\sum_{k}\int\int_{V_{ik}}\int_{z=0}^{q(p_{ik},r,t)} p(z, r, t)\,dz\,dt\,dr - C.$$

In the spatial problem, total demand was limited by capacity, while in (2) demand at each instant is limited to Q, so that in the mixed problem we simply combine these to yield

$$\sum_{i}\int_{U_i} q(p, r, t)\,dr \leqslant Q \quad \text{for all} \quad t \in [0, 1].$$

---

[16] In [4].

[17] In [1].

[18] In [1].

This form of constraint again ignores the possibility of forcing consumers to wait for the product, and also takes no account of the "outage" of equipment for maintenance. Each of these would involve uncertainties of demand or supply, and would reduce the capacity excesses which are characteristic of the temporal problem of Craven.[19] Once again, however, the distributional indifference implicit in the surplus maximizing approach is not suited to the analysis of such problems. The expression for surplus can be maximized subject to this constraint by choosing the p_{ik}, U_i, T_k, Q, I, and K, given a cost function C. This is a mathematical exercise whose form can be inferred from Craven[20] and earlier parts of this paper.

Even with a separable and constant returns cost function,

$$C = \sum_{i} \sum_{k} \int \int_{V_{ik}} g \cdot q(p_{ik}, r, t) dt dr + bQ,$$

there is no guarantee that the utility will break even when surplus is maximized, for this model will inherit the features of the model of the earlier Craven paper,[21] which, as we outlined above, prevented it from breaking even. Although an expression for the profit made by the utility can be derived, little can be said upon its sign or magnitude.

References

1. Craven, J. "On the Choice of Optimal Time Periods for a Surplus Maximizing Utility Subject to Fluctuating Demand." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 495–502.
2. Lancaster, K. *Mathematical Economics*, Chapter 5. New York: Macmillan, 1968.
3. Mohring, H. "The Peak Load Problem with Increasing Returns and Pricing Constraints." *The American Economic Review*, Vol. 60, No. 4 (September 1970), pp. 693–705.
4. Williamson, O. E. "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints." *The American Economic Review*, Vol. 56, No. 4 (September 1966), pp. 816–827.

---

[19] In [1].

[20] See [1].

[21] See [1].

Corporate diversification and market structure

Charles H. Berry

Professor of Economics and Public Affairs
Princeton University

Views differ regarding the structural implications of diversification by large industrial corporations. As yet, there is little definitive empirical evidence. This paper, piecing together data from a variety of sources, seeks to test the proposition that this diversification has tended to strengthen the market position of leading firms in already concentrated industries. It finds the opposite. The projected market share of leading firms is shown to be reduced by the entry of large corporations. The more concentrated the entered industry, the greater the reduction. The paper also demonstrates the presence of barriers insulating the market position of leading firms in concentrated industries from the competition of entering smaller firms.

1. Introduction

■ Writing in 1970, and basing his remarks largely on data from the late nineteen-fifties and early sixties, William G. Shepherd appraised the effect of increasing diversification by large corporations in the following manner:

> Very broadly speaking, diversification in the United States is associated with concentration. . . . One may first look at the incidence of the very largest firms in markets. In 264 product classes out of the total of 926 in 1950, at least one of the 1,000 largest firms was among the leading four producers. . . . In 1958 a majority (550) of all 1,014 product classes had a branch of at least one of the 100 largest companies among its four leading firms. But less than one-tenth (86) of the product classes had one or more such branches among its next four firms and at the same time, none in the top four. In 1963, the patterns were much the same. *Therefore, one may fairly conclude that branching by these firms tends to reinforce market power, rather than to neutralize it.* [Emphasis added.][1]

Summarizing his analysis, Shepherd himself concludes that the data he examines "confirm the reinforcing character of structural elements, and they go against the grain of the Schumpeterian view of entry and diversification. Creative destruction apparently destroys the hapless rather than the strong."[2]

This paper reports an alternative test of this thesis that this

---

Development of the data on which the empirical portions of this paper are based was supported by the Brookings Institution under a grant from the Alfred P. Sloan Foundation. The paper was prepared while the author was Visiting Professor at the University of California, San Diego, and an earlier version circulated as Discussion Paper 73-10, Department of Economics, U.C.S.D.

Comments and suggestions from Dennis E. Smallwood and Stephen M. Goldfeld, and from a not-so-anonymous referee (Jerry Hausman), are very gratefully acknowledged.

[1] In [7], pp. 140–141.

[2] *Ibid.*, pp. 143–144.

"branching" (entry) by large corporations to new industries has tended to reinforce market power (concentration). It finds exactly the reverse.

2. A simple model of structural change

■ The most common measure of market structure is, of course, the four-firm concentration ratio: the ratio of the value of shipments of the four largest firms within a given market to the total value of shipments of all firms within that market. To the extent that entry by large corporations can be shown either to "reinforce" market power or to erode it, one place to look for that effect would be in its impact on this measure of market structure. That, in essence, is the basis of Shepherd's test. It is also the basis of the test reported here.

Let $VS_1{}^4$ and $VS_2{}^4$ be the value of shipments of the four largest sellers within a given industry in the first and last years, respectively, of the time interval under consideration. Correspondingly, let VS_1 and VS_2 be the total value of shipments of the industry in those two years. The ratio of concentration in the final year to concentration in the initial year may be written as $\frac{VS_2{}^4}{VS_2} \cdot \frac{VS_1}{VS_1{}^4}$. Consider, however, the determinants of $VS_2{}^4$. If concentration were to remain unchanged, then:

$$VS_2{}^4 = \alpha_0 \frac{VS_2}{VS_1} \cdot VS_1{}^4, \qquad (1)$$

with $\alpha_0 = 1$ and with the four largest shippers' sharing proportionately in the growth or decline of the industry. If, other things being equal, industry growth fosters increasing concentration, $\alpha_0 > 1$; if such growth leads to decreasing concentration, $\alpha_0 < 1$.

Now let NC_1 and NC_2 be the total number of companies present in the industry in years 1 and 2, and consider:

$$VS_2{}^4 = \alpha_0 \left[1 + \alpha_1 \frac{NC_2 - NC_1}{NC_2} \right] \frac{VS_2}{VS_1} \cdot VS_1{}^4. \qquad (2)$$

If the number of companies is unchanged, (2) reduces to (1). The parameter α_1 in (2) simply adjusts the relationship in (1) for total entry: change in the total number of companies present. The expression is structured so that if α_1 were -1, the implication would be that the growth-adjusted value of shipments of the four largest sellers within the industry would vary in inverse proportion to the total number of companies present. That would be true if the size distribution of firms in the industry were independent of the number of firms present—that, other things being equal, a doubling of the number of firms would result in a halving of the average size of firms in each size class. If, however, the market position of leading firms were totally insulated from this "overall" entry, α_1 would be zero. In the unlikely event that such entry somehow facilitated growth by leading firms, α_1 would be positive.

Here, however, the issue is not this overall entry (or entry presumably by "small" corporations), but entry by corporations that are large diversifying firms. Let $VS_2{}^E$ be the total value of shipments within the industry in year 2 of any such entering *large* firms. Consider

$$VS_2^4 = \alpha_0 \left[1 + \alpha_1 \frac{NC_2 - NC_1}{NC_2} \right] \frac{VS_2}{VS_1} \cdot VS_1^4 + \alpha_2 VS_2^E. \quad (3)$$

If this entry (VS_2^E) were to reduce, on a dollar-for-dollar basis, the otherwise expected value of shipments of the leading four firms, α_2 would be -1. That would be an extreme illustration of the "creative destruction" Shepherd finds to be absent. The Shepherd thesis is that α_2 is nonnegative—that "branching by these (large) firms tends to reinforce market power ($\alpha_2 > 0$), rather than to neutralize it" ($\alpha_2 < 0$). Any negative value of α_2 between 0 and -1 would indicate the proportion of the total market position acquired by entering *large* firms (VS_2^E) that was obtained by an equal reduction in the market position of leading firms (VS_2^4). Note that an entering large firm may become a leading firm, and hence that a positive value for α_2 is not an implausible result.

This is the basic form of the test employed here. Since four-firm concentration in year 1 ($CON4_1$) is simply $\frac{VS_1^4}{VS_1}$ and four-firm concentration in year 2 ($CON4_2$) is $\frac{VS_2^4}{VS_2}$, (3) is readily rewritten as

$$\frac{CON4_2}{CON4_1} = \alpha_0 + \alpha_0\alpha_1 \frac{NC_2 - NC_1}{NC_2} + \alpha_2 \frac{PEN}{CON4_1}, \quad (4)$$

where $PEN = \frac{VS_2^E}{VS_2}$, or the market share (penetration) in year 2 of the entering large firms.

It is, however, reasonable to expect that the values of both α_1 and α_2 would be affected by the initial structure of the industry in question. For example, that:

$$\alpha_1 = \beta_0 + \beta_1 \, CON4_1 \quad (5)$$

and

$$\alpha_2 = \lambda_0 + \lambda_1 \, CON4_1 \quad (6)$$

In the case of α_2, if concentration were very low—implying an industry in which leading firms are relatively small—it is indeed plausible that an entering large firm might become a leading firm and that α_2 might be positive. But in the light of Shepherd's "creative destruction," that case is uninteresting: there is nothing to destruct creatively in the first place. The interesting question is the value of $\alpha_2 = \lambda_0 + \lambda_1 \, CON4_1$ where $CON4_1$ is high—where structural imperfection *is* present, and where entry by large firms *might* be capable of the competitive effect Shepherd denies.

Equally interesting is the parallel interpretation of β_0 and β_1 in the case of entry by "other" firms.[3] If the market position of leading firms is readily eroded by the entry of *any* firm, it is somewhat beside the point to worry about the entry of large firms. There ought to be enough other ("small") firms to go around. But it is entirely possible—as I have argued at length elsewhere—that the *important* structural imperfections arise where large leading firms in concentrated industries are able to protect by some "barrier" their market position from the competitive pressure of smaller firms—either smaller firms within the industry, or small firms that enter or attempt to enter that industry.[4] Empirical support for this sort of hypothesis would be found in a strongly *negative* value (approaching -1) for β_0 with an offsetting *positive* value for β_1, such that $\alpha_1 = \beta_0 + \beta_1\, CON4_1 \approx 0$ for highly concentrated industries. Under these circumstances, the value of α_2 would become doubly interesting. With evidence of leading firms' being immune to the entry or penetration of small firms in concentrated industries, a high negative value for λ_1 would suggest "creative destruction" by large entering firms precisely where it is needed. Correspondingly, a positive value for λ_1 would support the Shepherd position in spades.

---

[3] In what follows, "large" firms are defined as the very largest industrials, of which a very small number at most enter any industry during any

With this modification, and as fitted, (4) becomes:[5]

$$\frac{CON4_2}{CON4_1} = \alpha_0 \left[1 + (\beta_0 + \beta_1\, CON4_1) \frac{NC_2 - NC_1}{NC_2} \right] + (\lambda_0 + \lambda_1\, CON4_1) \frac{PEN}{CON4_1} + \mu \tag{7}$$

or

$$\frac{CON4_2}{CON4_1} = \alpha_0 + \alpha_0\beta_0 \frac{NC_2 - NC_1}{NC_2} + \alpha_0\beta_1 \frac{NC_2 - NC_1}{NC_2} \cdot CON4_1 + \lambda_0 \frac{PEN}{CON4_1} + \lambda_1\, PEN + \mu. \tag{8}$$

3. Some empirical results

■ This section makes use of data developed from the 1961 and the 1966 *Plant and Product Directory* compiled by *Fortune*.[6] Each directory lists the approximate employment and the five-digit products of each of the plants of largest industrial corporations. The earlier *Directory* includes 1960 information for 497 corporations in that year. The 1966 *Directory* covers the 1,000 largest industrials in 1965. It is possible to assemble from these data estimates of the 1960 and 1965 employment, by four-digit industries, for each of 461 large industrials.[7] Those corporations

---

given time period. These "other" firms are, therefore, in practice "small firms." Results and interpretation would not differ if the number of companies were defined exclusive of large firms.

[4] See Berry [2].

[5] Note that if (5) and (6) are considered stochastic, this specification is obviously heteroscedastic. Correction is not readily feasible. See Theil [8], pp. 622–627.

[6] [3].

[7] For additional detail with respect to these data and their processing, see Berry [1].

are the 500 largest (in terms of sales) in 1960, less three corporations that reported no domestic manufacturing activity in 1960 (and were excluded from the 1961 *Directory* on that basis), and less 36 corporations that merged out of existence (or were liquidated) between 1960 and 1965. These are taken as "large corporations" in the context of the foregoing discussion. The variable VS_2^E is defined for each four-digit industry as the total estimated employment, times $25,000, of all plants of all corporations among those 461 entering the industry between 1960 and 1965. The $25,000 is an approximation of output per worker in manufacturing establishments.[8] The variable *PEN* is this measure of entry by large corporations (VS_2^E) divided by the value of shipments in the industry in 1963.[9] Other variables in the model are value of shipments, number of companies, and four-firm concentration by four-digit industries. These measures are available from the Census of Manufacturers for 1958, 1963, and 1967.[10] It was possible to assemble all variables in each year for 200 four-digit industries.[11] The Census years do not, unfortunately, match those for which the large corporate entry variable can be defined. Table 1, therefore, shows regression results for change in concentration over three alternative time periods: 1958 to 1967, 1958 to 1963,

TABLE 1

REGRESSION COEFFICIENTS AND t-RATIOS CHANGE IN 4-FIRM CONCENTRATION, 1958-1967, 1958-1963 AND 1963-1967, REGRESSED ON SELECTED INDEPENDENT VARIABLES, 200 4-DIGIT MANUFACTURING INDUSTRIES[a]

REGRESSION COEFFICIENT	TIME PERIOD		
	1958-1967	1958-1963	1963-1967
$\hat{\alpha}_0$	1 067 (39 64)	1 047 (57 65)	1 031 (80 67)
$\hat{\beta}_0$	−0 745 (−5 22)	−0 539 (−3 22)	−0.671 (−4 97)
$\hat{\beta}_1$	1 067 (3 63)	0 722 (2 11)	0 966 (3 68)
$\hat{\lambda}_0$	0 526 (4 15)	0 351 (3 91)	0 093 (1 43)
$\hat{\lambda}_1$	−1 670 (−4 25)	−1 071 (−3 86)	−0 433 (−2 19)
R^2	0 24	0 14	0 17

[a] SEE TEXT, EQUATION (8) AND FOOTNOTE 12 THE NUMBERS IN PARENTHESES ARE t-RATIOS

---

[8] See, for example, [10]. Note that the effect of this correction is only to scale proportionately the estimated values (but not the significance) of λ_0, λ_1 and α_2. It is made to facilitate interpretation of those parameters.

[9] The data in [3] do not identify plants acquired by merger. *This measure of entry therefore includes "entry" by merger.* It is not possible with these data, or with any other such data for this period, to distinguish entry by merger from entry resulting from the creation of new facilities without any form of corporate acquisition. For the purposes that the data are put here, this is probably not a drawback. For large corporations, the conventional means of entry probably *is* merger.

[10] See [9].

[11] Parts of this analysis were repeated (though are not reported here) using product class concentration ratios. The industries considered include all for which these class ratios, as well as the other variables considered, are available for all years.

and 1963 to 1967. Parameters are defined in the table with the notation introduced earlier. In each case, the measure of large firm entry (*PEN*) is the same. All other variables are specific to the time period indicated. In each case as well, the fit is to observations for the same 200 four-digit industries.[12]

In Table 1, $\hat{\alpha}_0$ is generally close to the expected (neutral) value of 1.0 and $\hat{\beta}_0$ is negative, again as expected. However, $\hat{\beta}_1$ is strongly positive: as concentration increases, the concentration-reducing effect of total entry *decreases*. The effect of entry by large firms, however, is almost exactly the opposite: $\hat{\lambda}_0$ is *positive*. When concentration is low, $\hat{\alpha}_2 \approx \hat{\lambda}_0$ and the effect of entry by large firms, other things being equal, is projected as an *increase* in concentration. But $\hat{\lambda}_1$ is strongly negative. As initial concentration is increased, $\hat{\alpha}_1$ also becomes negative. For convenience, Table 2 contains the values of $\hat{\alpha}_1 = \hat{\beta}_0 + \hat{\beta}_1\ CON4_1$ and of $\hat{\alpha}_2 = \hat{\lambda}_0 + \hat{\lambda}_1\ CON4_1$ implied for different levels of concentration by the regression coefficients of Table 1.[13] Table 2 shows the clear ten-

TABLE 2

CALCULATED VALUES[a] OF $\hat{\alpha}_1 = \hat{\beta}_0 + \hat{\beta}_1 CON4_1$ AND OF $\hat{\alpha}_2 = \hat{\lambda}_0 + \hat{\lambda}_1 CON4_1$ FOR ILLUSTRATIVE VALUES OF $CON4_1$

CON 4_1	TIME PERIOD					
	1958-1967		1958-1963		1963-1967	
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
0.25	−0.476 (−5.64)	0.109 (1.41)	−0.359 (−3.78)	0.083 (1.52)	−0.430 (−5.39)	−0.015 (−0.40)
0.40	−0.314 (−4.73)	−0.142 (−1.48)	−0.250 (−3.70)	−0.078 (−1.16)	−0.285 (−4.89)	−0.080 (−1.76)
0.60	−0.099 (−1.19)	−0.476 (−3.06)	−0.106 (−1.27)	−0.293 (−2.67)	−0.092 (−1.41)	−0.166 (−2.21)
0.75	0.062 (0.53)	−0.725 (−3.48)	0.002 (0.02)	−0.453 (−3.08)	0.053 (0.57)	−0.231 (−2.27)

[a] THE NUMBERS IN PARENTHESES ARE t-RATIOS NOTE THAT STRICTLY SPEAKING, THESE ESTIMATES ARE NOT DISTRIBUTED AS t SEE TEXT, FOOTNOTE 13

---

[12] The parameters of interest in (8) above—e.g., the vector $\theta = (\alpha_0, \beta_0, \beta_1\ \lambda_0, \lambda_1)$—can be estimated either by ordinary least squares or by nonlinear techniques. Point estimates will be identical in the two cases. However, nonlinear techniques permit calculation of standard errors for all components of θ, whereas, with the specification of (8), ordinary least squares does not. Appropriate variances and covariances were therefore obtained by inverting

$$R = \frac{1}{2\hat{\sigma}^2}\left[\frac{\partial^2 S}{\partial\theta\partial\theta'}\right],$$

where $\hat{\sigma}^2$ is the estimated residual variance and S the sum of squares. This is the equivalent of using the Cramer-Rao bound in the context of maximum likelihood estimation. See Goldfeld and Quandt [4], Chapter 2.

[13] Standard errors for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ were calculated from the variance-covariance matrices for the parameters of θ. See above, note 12. As a check on the accuracy of this procedure, exact confidence intervals for $\hat{\alpha}_1$ were also calculated by solving the quadratic in α_1 of:

$$(\hat{\alpha}_0\alpha_1 - \widehat{\alpha_0\beta_0} - \widehat{\alpha_0\beta_1}CON4_1)^2 = t^2\hat{s}_z^2,$$

where $z = \hat{\alpha}_0 - \widehat{\alpha_0\beta_0} - \widehat{\alpha_0\beta_1}\ CON4_1$, t is the appropriate t-statistic, and $\hat{s}_z^2$

dency for the estimated effect of entry by large firms ($\hat{\alpha}_2$) to become strongly negative as the level of concentration increases. It also shows the reverse tendency for the calculated value of $\hat{\alpha}_1$ (the coefficient on total or "small" firm entry) to approach zero as concentration increases. Those results are consistent with the thesis that where concentration is high, the market position of leading firms tends to be protected from erosion through the entry of "small" firms ($\hat{\alpha}_1$), but not from entry by firms among the 461 largest ($\hat{\alpha}_2$). Furthermore, the estimated value of α_2 is such that the projected impact is not small. A coefficient of —0.7 implies that 70 percent of the market share of entering large firms is acquired at the expense of the leading four firms. It would be difficult to imagine results that would more dramatically contradict the Shepherd hypothesis.

4. Changing eight-firm concentration

■ Tables 3 and 4 show corresponding results for change in eight-firm concentration where the fitted form is

$$\frac{CON8_2}{CON8_1} = \alpha_0 + \alpha_0\beta_0 \frac{NC_2 - NC_1}{NC_2} + \alpha_0\beta_1 \frac{NC_2 - NC_1}{NC_2} \cdot CON8_1 + \lambda_0 \frac{PEN}{CON8_1} + \lambda_1 PEN + \mu, \tag{9}$$

where *CON*8 is the eight-firm concentration ratio with, as before, the subscripts indicating the initial or the terminal year of the time period in question.[14]

TABLE 3

REGRESSION COEFFICIENTS AND t-RATIOS CHANGE IN 8-FIRM CONCENTRATION, 1958-1967, 1958-1963 AND 1963-1967, REGRESSED ON SELECTED INDEPENDENT VARIABLES, 200 4-DIGIT MANUFACTURING INDUSTRIES[a]

REGRESSION COEFFICIENT	TIME PERIOD		
	1958-1967	1958-1963	1963-1967
$\hat{\alpha}_0$	1 054 (53 89)	1 038 (77 03)	1 022 (111 99)
$\hat{\beta}_0$	−0 668 (−5 69)	−0 528 (−3 62)	− 0 735 (−6 55)
$\hat{\beta}_1$	0 744 (3 89)	0 550 (2 42)	0 876 (5 13)
$\hat{\lambda}_0$	0 727 (4 74)	0 237 (3 56)	0 242 (3 21)
$\hat{\lambda}_1$	−1 571 (−4 86)	−0 790 (−3 85)	−0 523 (−3 32)
R^2	0 28	0 15	0 27

[a] SEE TEXT, EQUATION (9) AND FOOTNOTE 14 THE NUMBERS IN PARENTHESES ARE t-RATIOS

---

is the estimated variance of z. Variances and covariances for $\hat{\alpha}_0$, $\widehat{\alpha_0\beta_0}$, and $\widehat{\alpha_0\beta_1}$ were in this case estimated by ordinary least squares. With t set at 1.96, virtually identical results were obtained with the two procedures. For a similar application of this technique, which is an old one, see Griliches [5].

[14] As with (8), standard errors for the parameters of interest were obtained through nonlinear estimation. See above, note 12.

TABLE 4

CALCULATED VALUES[a] OF $\hat{\alpha}_1 = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{CON8}_1$ AND FOR $\hat{\alpha}_2 = \hat{\lambda}_0 + \hat{\lambda}_1 \mathrm{CON8}_1$ FOR ILLUSTRATIVE VALUES OF $\mathrm{CON8}_1$

$\mathrm{CON8}_1$	TIME PERIOD					
	1958-1967		1958-1963		1963-1967	
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
0.25	−0.482 (−6.31)	0.333 (3.49)	−0.391 (−4.14)	0.040 (0.99)	−0.516 (−6.98)	0.109 (2.39)
0.40	−0.370 (−6.48)	0.098 (1.19)	−0.308 (−4.56)	−0.079 (−1.58)	−0.385 (−7.10)	0.029 (0.76)
0.60	−0.221 (−4.51)	−0.216 (−2.06)	−0.199 (−4.03)	−0.237 (−2.92)	−0.210 (−5.25)	−0.078 (−1.57)
0.75	−0.110 (−1.81)	−0.451 (−3.23)	−0.115 (−1.95)	−0.355 (−3.26)	−0.078 (−1.70)	−0.157 (−2.34)

[a] THE NUMBERS IN PARENTHESES ARE t-RATIOS SEE TEXT, FOOTNOTE 13

The regression results of Table 3 closely follow those of Table 1. Some differences between the two tables are to be expected. For example, the estimated coefficient $\hat{\lambda}_0$ is larger for change in eight-firm concentration than for change in four-firm concentration: where concentration is low, it is *a priori* more likely that an entering large firm would displace one of the eight leading firms than one of the top four. That is what the comparison indicates. Similarly, the estimated coefficient $\hat{\beta}_1$ is lower in Table 3 than in Table 1: where concentration is high, it is (perhaps) more likely that entry by "small" firms would reduce the market share of the leading eight than that of the leading four. Even these differences, however, are readily exaggerated, since eight-firm concentration will always exceed four-firm concentration. Taking that into account, the two sets of results are strikingly similar, though they are, of course, scarcely independent. Nevertheless, repeating the test with market shares of the eight leading firms does not alter the general conclusion: "branching" by these 461 large firms during this period does *not* appear by this test to have been a factor tending to reinforce concentration. The data suggest quite the opposite.

5. A caveat

■ Having chosen to report these results within the context of the contrary conclusions of a close friend and good colleague, it is certainly appropriate to note that Shepherd conditions his conclusions with the observation:

> Accurate and full evidence . . . would not be difficult for the Census Bureau to compute, and it would go far to clarify the real causes and effects of diversification.[15]

That is also the conclusion of this paper. Given the tortuous (and indeed torturing) exercise in the development of what are most certainly the very crude measures of diversifying activity employed here, that the results reported are as consistent as they are, is (1) little short of astonishing, and (2) illustrative of the need for those more accurate and hopefully more revealing Census tabulations.

---

[15] Shepherd [7], p. 140.

For example, the most recent major proposal for antitrust reform calls for merger rules severely restricting acquisitions by large firms with the language:

> Our proposal focuses on the second probability, that a substantial number of large corporations will enter a substantial number of concentrated industries, *and is intended to channel the potential competition of large firms along lines that are conducive to reducing levels of concentration in the American economy.* [Emphasis added.][16]

It is exactly this "channeling" that the Task Force fails to find, that this paper clearly detects, and that Shepherd's Census tabulations might or might not reveal. It would be nice to know.

Quite apart from that, however, the most interesting and worrisome result is the relatively strong unanticipated evidence that the market position of leading firms *does* appear protected in concentrated industries from entry by other than large firms. That is the bad news. The good news is that at least some of the diversifying activity of large firms *may* have filled the gap.

References

1. Berry, C. H. "Corporate Bigness and Diversification in Manufacturing." *Ohio State Law Journal*, Vol. 28, No. 3 (Summer 1967), pp. 402–426.
2. ———. "Economic Policy and the Conglomerate Merger." *St. John's Law Review*, Vol. 44, Special Edition (Spring 1970), pp. 266–281.
3. Fortune Magazine, Market Research Department. *Plant and Product Directory*. New York: Time, Inc., 1961 and 1966.
4. Goldfeld, S. M. and Quandt, R. E. *Nonlinear Methods in Econometrics*. Amsterdam: North Holland, 1972.
5. Griliches, Z. "Distributed Lags: A Survey." *Econometrica*, Vol. 35, No. 1 (January 1967), pp. 32–33.
6. *Report of the White House Task Force on Antitrust Policy*. Washington, D. C., 1969.
7. Shepherd, W. G. *Market Power and Economic Welfare*. New York: Random House, 1970.
8. Theil, H. *Principles of Econometrics*. New York: John Wiley, 1971.
9. U. S. Department of Commerce, Bureau of the Census. *Concentration Ratios in Manufacturing Industry, 1967*. Washington, D. C.: U. S. Government Printing Office, 1971.
10. U. S. Department of Commerce, Business and Defense Services Administration. *Industry Profiles, 1958–1967*. Washington, D. C.: U. S. Government Printing Office, 1969.

---

[16] See [6], p. III-3.

The CAB's domestic passenger fare investigation

George W. Douglas

Associate Professor of Economics
University of North Carolina at Chapel Hill

and

James C. Miller III

Associate Professor of Economics
Texas A&M University

For the second time in its 35-year history, the CAB has undertaken a general, in-depth investigation of airline pricing policies. This Domestic Passenger Fare Investigation is described below, primarily from the standpoint of the presentations made by the U. S. Department of Transportation, for whom the authors served as expert witnesses. The investigation is important in two specific respects: first, it marks what could prove to be a significant and possibly "enlightened" change in CAB regulatory policy; and second, the analysis presented in the case should be applicable to a broad range of quasi-competitive industries subject to economic regulation. But while the decisions rendered in the investigation form grounds for optimism, recent Board actions raise severe questions about the permanency of this new policy direction.

1. Introduction

■ Just as in civil and criminal law, the importance of decisions in individual regulatory cases in molding agency policy should not be underestimated. Since the Civil Aeronautics Board (CAB or Board) is loath to have its decisions overturned in the courts, it is biased toward rendering judgments in individual cases consistent with earlier ones which were successful in meeting court tests. Occasionally, however, the Board will institute an investigation of a particularly troublesome issue and will call into question prior precedent. The Board feels confident in taking this sort of initiative since a precedent-breaking decision is unlikely to be overturned so long as it is reached upon a full and impartial hearing of the evidence (and, of course, so long as the decision and procedures used are consistent with the governing statutes).[1]

A preliminary version of this paper was presented by Dr. Miller at an AT&T Seminar on Problems of Regulation and Public Utilities at the Amos Tuck School of Business Administration of Dartmouth College on August 30, 1971. Helpful comments by George C. Eads and William A. Kutzke, and support from the Brookings Institution and the Texas A & M Research Foundation are gratefully acknowledged.

[1] In such a manner the Board may be relatively confident of successfully defending precedent-breaking decisions whether the rationale for change be an admission of error in the (formerly) precedent decisions,

The CAB's most important case of this type, where prior precedent has been brought into question, is the recent (nearly completed) Domestic Passenger Fare Investigation (DPFI).[2] In it Board policy in nearly every area of fare regulation was open to reassessment. For several reasons this case is worth summarizing and bringing to the attention of regulatory economists and practitioners. First, it indicates a new direction for CAB fare policy. Second, it shows how the direction of CAB policy can be changed and on what grounds such change can be justified. Third, it provides a case study of how economic analysis, forcefully advocated by another government agency (e.g., the U. S. Department of Transportation) may have an impact on a regulatory agency. Finally, the model of industry behavior discussed in the case and ultimately adopted by the Board should have broad application to other quasi-competitive industries under economic regulation.

A matter of perspective should be noted at the outset. The authors participated in the DPFI as expert witnesses for the U. S. Department of Transportation (DOT), and in that capacity presented materials and advocated positions consistent with maximizing economic efficiency, *given* the regulatory framework and given the scope of the proceeding.[3] Since our presentation constituted the leading "unconventional" view, we emphasize it in the discussion to follow.[4] Second, as will become apparent, the authors take some pride in having been associated with the "winning" side on most major issues. Whether in general "good" economics won out over "bad" because of luck, underlying forces, relative political clout of participants, the persuasive powers of the authors and others who took similar positions, or for some other reason(s), we cannot assess. However, because of this relative success, there may be a tendency to overstate what we have perceived as a new, relatively "enlightened" policy direction. It must be kept in mind that just as a plaintiff can and usually will sue on all grounds he thinks relevant,[5] the Board will attempt to justify its decisions with a multitude of reasons.[6] Consequently, each of the decisions rendered in this case expresses numerous justifications in addition to

---

or that the (former) precedent decisions no longer apply because the underlying conditions have changed.

[2] CAB Docket 21866; instituted by CAB Order 70-1-147 (January 29, 1970). As of this writing, the Board has yet to render a final decision in the last of nine "phases" of the investigation. Decisions in all previous phases were rendered during 1971 and 1972.

[3] The standard disclaimer: the authors do not necessarily endorse all the presentations made by the Department in that proceeding, nor, of course, should the authors' (unofficial) writings be interpreted as representing Departmental policy.

[4] For a summary of the positions of the various parties in the case, see Miller [3]

[5] For example, an airline may petition the Board to suspend or find unlawful a competitor's fare reduction under Sections 404(b)(Discrimination), 411 ([Unfair] Methods of Competition), and 1002(e)(Reasonableness, etc.) of the Federal Aviation Act.

[6] In other words, the Board will give reasons *A*, *B*, and *C* for a decision, even though each may appear to be separately sufficient. The reason is that in the event the decision is challenged in court, the hope is that at least one of the justifications will hold up.

those we cite. As the decisions pertain to economic analysis, however, they are generally consistent with essential points summarized.

A related issue is the permanency of this new policy direction. Since Board decisions are rendered in closed sessions, the public does not necessarily learn the *real* reasons for the positions chosen. Since professional staff members draft the decisions and since their job is to write a "tight" opinion that will stand up in court, the underlying rationale may never show through. In short, it is hazardous to impute to Board members the same decision calculus as expressed in the formal decision. Thus, while it might appear that the Board has moved toward a policy more consistent with economic efficiency, this change may have been simply expedient. If so, then if certain conditions change, then so may Board policy. In point of fact, as we shall describe later, one observes already the Board's retreating from the DPFI findings.

2. Nature of the DPFI

■ The DPFI is the Board's second, general, in-depth investigation of airline pricing policies. The first was held over the period 1956 to 1960, and became known as the General Passenger Fare Investigation (GPFI). It was brought on principally by Congressional pressure to see that the Board correctly discharged its rate-making responsibilities, by carrier filings for fare increases during the 1951 to 1956 period, and by the Board's need to determine what kind of fare increase (if any) might be warranted and what standards should be adopted to judge future fare filings.

Although considerable resources were devoted to the GPFI, the results were disappointing inasmuch as little was decided. In simplified form, the final conclusions were: (1) the reasonable rate of return on investment in commercial aviation is 10.5 percent,[7] and (2) the overall fare level should reflect the costs the carriers actually incur in providing service.[8] These limited findings subsequently circumscribed CAB fare policy throughout the decade of the 1960s.[9]

What brought on the DPFI was almost surely the Moss Case of 1970.[10] When the airlines' profits began falling in 1969, the Board, concerned with the carriers' plight, began holding *ex parte* informal meetings to work out alternative solutions. A group of Congressmen, headed by John E. Moss (D-California), complained of these actions and requested at minimum permission to attend. This petition was rejected. Subsequently, the carriers filed new fare increases. Although the Board rejected this proposal, it in turn indicated a fare (-mileage) formula that would be acceptable. The carriers then filed the Board-suggested formula.

---

[7] After taxes, but before interest payment on debt. The rate base was taken to be equity plus debt. An operating ratio approach to reasonable return was explicitly rejected.

[8] An attempt by the Board's Bureau of Air Operations to incorporate load factor standards was contemplated but rejected.

[9] For a thorough review of the GPFI, see Redford [4]. The Board's final decision in the case is contained in Order E-16068 (September 25, 1960).

[10] *John E. Moss et al. v. Civil Aeronautics Board.* For a summary of this case, see Judge J. Skelly Wright's opinion in the unanimous decision rendered in the U S. Court of Appeals for the District of Columbia Circuit, No. 23,627, decided July 9, 1970, reprinted in the *Congressional Record* (July 9, 1970), pp. H6569-73.

The Moss petitioners carried this action to court, arguing that in its haste to be responsive to the carriers, the Board had ignored certain statutory requirements. Partly in an effort to relieve public pressure over this case and partly, perhaps, to affect the outcome of the court's decision, the Board announced the DPFI on January 29, 1970, citing as reason "an exploration of the issues raised by the [Moss Group]."[11]

In its written opinion, the Court scolded the Board for its "blatant attempt to subvert the [Federal Aviation Act]" and for an "intimation . . . that its responsibilities to the carriers are more important than its responsibilities to the public."[12] The Court further declared the fares that had been approved by the Board unlawful (because of procedure) and remanded the case to the Board for further proceedings. The Board then merged certain of the court-raised issues into the DPFI, and set down a separate investigation to determine the lawfulness of the fares then in effect.[13]

As set forth in the order defining the scope of the DPFI, there were to be nine separate "phases":[14]

(1) Aircraft Depreciation
(2) Leased Aircraft
(3) Deferred Federal Income Taxes
(4) Joint Fares
(5) Discount Fares
(6) Load Factor and Seating Configurations[15]
(7) Fare Level
(8) Rate of Return
(9) Fare Structure.

The first three phases would be handled by rule-making proceedings, with the remaining six set down for public hearing.[16]

Cases involving airline fares are, by convention, broken down into issues involving fare *level* (or average yield) and fare *structure* (or the relationship one fare has to another). The fare level issues in the DPFI are contained primarily in Phases 1, 2, 3, 6, 7, and 8. Fare structure issues are the principal subjects of Phases 4, 5, and 9.

---

[11] CAB Order 70-1-147.

[12] *Congressional Record*, p. H-6571.

[13] CAB Docket 23140. The Board has since determined that the fares charged during this period were, in fact, unlawful. See CAB Order 73-7-39 (July 13, 1973).

[14] CAB Order 70-2-121 (February 26, 1970).

[15] By Order 70-11-91 (November 19, 1970) the Board separated Phase 6 into Phase 6A-Seating Configurations, and Phase 6B-Load Factor.

[16] Rule-making proceedings involve only exchange of written communication. Public hearings involve cross-examination of witnesses before an Administrative Law Judge (formerly "Hearing Examiner"), and usually, though not always, oral argument before the full Board.

3. Fare level issues[17]

Essentially the Board's responsibility in the fare level phases of the DPFI was to decide between two opposing regulatory philosophies. The more traditional, industry view held that the Board's function is that of determining the actual (reasonable) costs of performing service and then regulating fares so as to cover these costs. This meant, in essence, observing whether the carriers' rate of return was excessive or insufficient, and then, based in part on predicted costs and revenues, regulating fare levels (taking into account demand price elasticity) in such a way as to match total revenues with total costs (including a reasonable return on investment).[18] The proponents of this position based their case primarily on Section 1002(e)(5) of the Federal Aviation Act, whereby in setting rates the Board must take into consideration, "the need of each air carrier for revenue sufficient to enable such air carrier, under honest, economical, and efficient management, to provide adequate and efficient air carrier service." Oversimplifying a bit, the carriers argued that unless the Board could prove dishonesty or gross inefficiency, it had to accept carrier costs as being justified, and therefore had no choice but to adjust revenues to cover the costs of whatever levels (and qualities) of output the carriers anticipated providing.

The alternative view held that the Board need not accept carrier costs as sole evidence of a reasonable fare level. Its advocates took several different perspectives. Some simply argued that "the public should not be made to pay for excess capacity," as evidenced by aircraft flying on average less than half-filled. Some suggested that carriers be penalized for flying excess schedules by not counting in their rate base the costs of flights averaging less than some X-percent load factor. The Department of Transportation argued that load factor and service quality are, in fact, endogenous to the market equilibrium; that once price is determined, nonprice competition (primarily in the form of schedule frequency) results in a (zero-rent) market equilibrium quality of service. Moreover, the Department implied that the Board's preoccupation with determining and aggregating the carriers' costs and with adjusting fares (taking into account demand price elasticity) in order to match revenues with these estimated costs not only constitutes a naive and basically incorrect interpretation of airline industry competitive dynamics, but overshadows the price/quality trade-off which should be of crucial concern. In short, the Department argued that the appropriate fare level depends on an estimate of the optimal feasible fare/quality combination.

The analysis we presented on behalf of the Department of Transportation is straightforward, but deserves some elaboration.[19]

---

[17] In this overview we shall skip Phases 1, 2, and 3. For a summary of the decisions in these phases, see Miller [3]. Also, see Table 1 *infra.*

[18] The Board's existing policy toward fare level regulation may also be characterized as attempting to match revenues with predicted costs (including a reasonable return on investment). (Cf., Order E-16068, *ibid.*) Although the Board retained the right to "disallow" certain costs from inclusion in the rate base, it never went so far as to attempt to regulate the level of output and the overall quality of service.

[19] For a complete summary and further development of our analysis, see Douglas and Miller [2].

FIGURE 1

INVERSE RELATIONSHIP BETWEEN AVERAGE FARE (COST) AND AVERAGE LEVEL OF CAPACITY UTILIZATION (LOAD FACTOR)

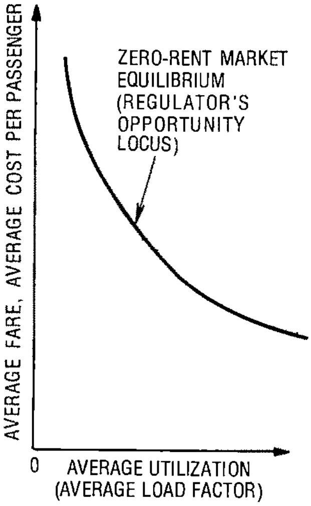
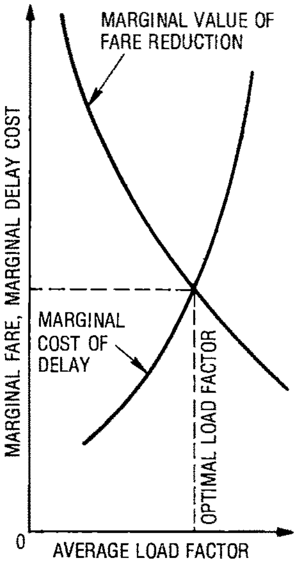


FIGURE 2

DETERMINATION OF OPTIMAL FARE/LOAD FACTOR COMBINATION



Essentially, we characterized the trunkline industr restrained cartel wherein carriers are seen to take pı lator-imposed parameter and "compete" (or rival) quality dimensions, primarily the frequency of sch cause of diminishing returns in the generation of n through adding flights, the profit-maximizing carrier extent of such activity, depending on price, the el mand, and other factors. Moreover, in markets wl several carriers (or where any carrier anticipates th itors will retain frequency at given levels or wh competition is expected to have a disproportionately impact on market share), profits will tend to be elimi of the inverse relation between average cost and ca tion (Figure 1). The empirical evidence would appe with this model (at least more so than with comp the industry).

An important conclusion which can be drawn f ysis is that the nature of the market equilibrium the regulator's choice of a price parameter. In othe is a broad range of alternative combinations of price utilization which are consistent with a zero-rent m rium (Figure 1). But note that the alternative cho ticular concern to the air traveller, since his "full cos depends not only on the ticket price but also on the cost involved in the service which, in turn, is a fı rate of capacity utilization. Technically, there is a d ship beween price and service quality, or more s inverse relationship between price and expected time for any given market, at a relatively high price, mark will be characterized by high schedule frequency factors; thus, frequency delay (i.e., the expected val interval between the desired time of departure and uled departure), stochastic delay (i.e., the expected the possibility of being unable to obtain accommod desired and subsequent scheduled departures), an (i.e., the sum of frequency and stochastic delays) wil small. At a low fare, the market equilibrium will be by a much greater expected total delay. Dependin senger's value of time (and other factors), there is a combination which minimizes his total trip cost. (analysis (illustrated for a single market in Figure 2) gested that the regulator's chosen level and structu inefficient, in that another set of alternatives woul lower passenger trip costs, yet yield normal profits t Specifically, we argued that optimal load factors ris distance, ranging in level between approximately 48 at short distances to approximately 62 to 67 percen tinental markets (actual levels depending on marke

---

[20] Section 401(e)(4) of the Federal Aviation Act prol from regulating carrier schedule frequencies. Under Sectio ever, the Board may (and on occasion has) approve(d) ca to "self-regulate." (See Section 7 *infra.*)

presumed value passengers place on total delay).[21] These estimated optimal levels were in contrast with the then-prevailing load factors which ranged from approximately 60 percent at short distances down to approximately 40 percent in the transcontinental markets. Thus, we argued, not only are the prevailing load factors "too low," but their distance structure is actually perverse (i.e., falling instead of rising with distance).

In the end, the Board accepted our nonprice competing cartel characterization of the industry, stating in its Phase 6 opinion:

> We find, as DOT has stated, that the higher the fare level in relation to cost, the more capacity carriers will offer and the lower load factors will be; and, conversely, the lower the fare level, the less capacity carriers will operate and the higher load factors will be.[22]

> In any given market, the carrier with the greatest number of schedules will normally carry the largest number of passengers. Thus, the desire to maximize market participation creates powerful incentives to add capacity. The countervailing incentive is supplied only by the imperative of economics: Schedules cannot be added indefinitely if the load factors achieved are insufficient, at the prevailing fare levels, to permit the carriers to cover costs and return a profit.[23]

This decision, which reversed the Administrative Law Judge's recommended opinion, clearly breaks with tradition.

Of course, a load factor standard must presume something about aircraft seating density. The Board's preliminary decision on this issue did not employ the nonprice competing cartel model and basically was inconsistent with the decisions rendered in other phases. Rather than basing fares on a standard load factor which, in turn, assumes some representative seating density, the Board announced that if carriers wished to configure their aircraft at lower than "standard" seating densities, they would have to collect minimum surcharges.[24] In its final order on the seating question, however, the Board changed its mind. It reverted to the cartel model, saying that it would base fares on "standard" seating densities and would not require surcharges.[25] Interestingly, the Board also stated that it would be receptive to proposals for fare *reductions* by carriers which chose to remove their coach lounges and utilize the standard (higher-density) seating.[26] Under this provision TWA was successful in reducing its one-way coach fare between New York and Los Angeles by $10 for aircraft not having coach lounges.[27]

---

[21] Cf. Douglas [1].

[22] CAB Order 71-4-54 (April 9, 1971), p. 23.

[23] *Ibid.*, p. 5.

[24] CAB Order 71-4-48 (April 8, 1971).

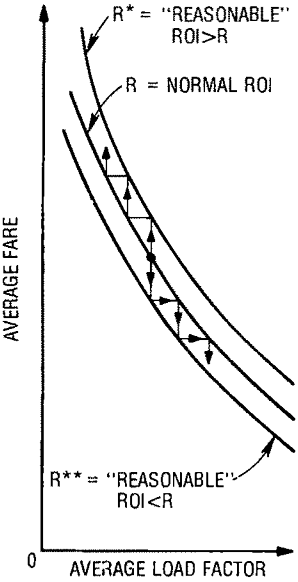
[25] CAB Order 72-5-101 (May 26, 1972).

[26] The petitioning carrier also had to prove that, as a result of utilizing the higher-density configuration, it was sustaining a significant adverse competitive impact from carriers which retained their lounges and charged the established fare.

[27] See CAB Order 73-1-69 (January 23, 1973). Also see CAB Order 73-6-102 (June 26, 1973) clarifying the Board's position on such fare differentials. It might be noted that Continental Airlines has taken this case to court, arguing that it should be allowed to match the $10 fare reduction and at the same time retain its coach lounges.

FIGURE 3

"RATCHET EFFECT" INSTABILITY WHEN "REASONABLE" RATE OF RETURN DIFFERS FROM NORMAL RATE



In its decision on rate of return the Board similarly accepted the nonprice competing cartel model and took due notice of the potential destabilizing effects of a policy of attempting to regulate rate of return precisely through changes in fares. That is, there is the potential that ignoring the interrelationships among fares, load factors, and profits can set up a dynamic between the industry and its regulators that *causes* instability. Suppose we observe an opportunity locus (or equivalently, an industry "reaction" curve) such as *R* in Figure 3. The location of this curve, of course, is predicated on the costs of production and the market rate of return (or cost of capital) as perceived by the firms. If, however, consistent with *R**, the CAB-defined "reasonable" rate of return exceeded the cost of capital as perceived by the carriers (or if the accounting conventions employed by the CAB had a consistent downward bias in measuring the carriers' earnings), then one could imagine a "ratchet effect" of regulation and reaction, in which price increases thought by the Board as necessary to raise profits only resulted in new equilibria with succeeding lower levels of capacity utilization. The ratchet effect would work in the reverse direction if the Board's computed reasonable rate of return, *R***, were less than the industry's cost of capital, *R*. Attempts to lower industry profits by reducing price would result in successive increases in capacity utilization (and reductions in service quality).

Thus, much less important than the rate of return actually deemed reasonable (the Board settled on 12 percent[28]) is the manner in which the rate is to be regulated.[29] The Board realized the futility of attempting to regulate rate of return precisely and in essence decided to use the calculated reasonable rate only as an aid in imputing a normal profit "cost" to be covered by a reasonable fare level. To quote the Board,

> The rates of the return specified herein will be used as standards for measuring the reasonableness of the general domestic passenger-fare level. They are not in any sense to be regarded as guarantees that any individual carrier will earn the standard return in any given year or period of years, or that the industry as a whole will achieve the specified rates in particular periods.[30]

Furthermore, in its fare level decision, the Board said that,

> The fact that earnings in a particular year are either above or below the standard rate of return would not be an occasion for fare adjustments unless the fares were significantly out of line with those required to produce

---

[28] See CAB Order 71-4-58 (April 9, 1971). The figure of 12 percent appears to overestimate the carriers' actual cost of capital inasmuch as in making this assessment the Board utilized an "optimal" debt/equity ratio of 45/55 rather than the actual ratio at 60/40. (CAB Order 71-4-58, *ibid.*) Based on estimated costs of debt (6.20 percent) and equity (16.75 percent), using the actual debt/equity ratio would imply a reasonable return of only 10.42 percent.

[29] Another reason we considered the actual rate chosen relatively unimportant is that since the airlines' yearly-revenue/total-investment ratio exceeds unity, a given (percentage) change in the estimated reasonable rate of return would mean a smaller (percentage) change in the cost-recovery level of fares.

[30] CAB Order 71-4-58, p. 3.

> reasonable earnings at a standard load factor. The Board believes that a firm adherence to this policy will enable the industry, during representative periods, to cover its cost of capital and provide investors with reasonable compensation for the risks taken and permit the carriers to attract needed additional capital. [Footnotes omitted.][31]

As for the "ratchet effect," the Board stated in its load factor decision:

> Increasing the fare level for the purpose of achieving profitable operations at the lower load factor lowers the break-even load factor and thus encourages the addition of more capacity, leading again to lower profits and demands for further fare increases. It is obvious that the actual capacity scheduled by the carriers is not the proper basis for determining reasonable future fares and that it is the duty of the Board to establish reasonable load-factor standards in order to discourage the uneconomic cycle of excess capacity and higher fares.[32]

While it is clear that in the fare level issues the Board implicitly accepted the nonprice competing cartel model described earlier, in selecting the specific load factor standard it neither endorsed our model of optimal price/quality estimation, nor did it approach the question in any sophisticated way. In fact, the decisions contain little justification for the standard chosen:

> For any fare-level adjustment that may be necessary at the conclusion of the *Fare Structure* phase, we intend to apply an overall trunk load-factor standard of 54 percent.[33]

This standard not only differs from our estimates of optimal average load factors, but it is in strong contrast with the price-competitive equilibrium load factors found in the Intrastate California markets, ranging generally between 65 and 75 percent.

It is possible, of course, that the yet to be rendered Phase 9 decision will contain an explicit justification for the chosen price/quality combination and/or will set forth a methodology for making such choices. Conceivably the Board might endorse outright the approach we advocated. In any event, it would appear that while the Board now recognizes that trade-offs between service quality and price do exist, it has yet to face deliberately the issue of choosing the optimal level and structure of price/quality configuration. As it stands, the decision changes little of substance, since the standard chosen is significantly short of reasonable estimates of optimal price/quality values.

---

[31] CAB Orders 71-4-59 and 71-4-60 (April 9, 1971), pp. 73 & 74.

[32] CAB Order 71-4-54, p. 21.

[33] CAB Orders 71-4-59 and 71-4-60, p. 50. The Board selected this particular standard primarily on the basis of calculations provided by the Bureau of Economics. In those, monopoly routes were presumed to attain load factors of 62 percent, and then an additional aircraft of average capacity (for that route) was added for each additional carrier in competitive markets. (No adjustment was made for the traffic-generating effect of additional flights and lower load factors.) Under these limiting assumptions, this resulted in an average industry load factor of 54 percent. The 62 percent figure was argued to be feasible (having been attained during a number of preceding years) and yet not so high as to turn away too much traffic in monopoly markets. The target of 54 percent was argued (subjectively) to strike a balance between attainable economies of passenger/aircraft density and the availability of service.

4. Fare structure issues

■ Regarding matters of fare structure, consistent with economic efficiency, we advocated that fares in individual markets approximate as closely as possible their respective marginal and average costs.[34] This position was in conflict with the Board and industry's traditional policy of (third-degree) price discriminating to augment traffic and total sales.[35] For example, fares in short-haul, relatively more price elastic markets were depressed artificially below costs, with excess profits from relatively less elastic long-haul markets making up the difference. Also, youth, families, vacation travellers, and others with relatively elastic demands were given special discounts, thus paying fares below those with less elastic demands (e.g., business travellers).

In the order instituting the investigation, the Board appeared to endorse a policy of efficient pricing in saying,

> A fare structure that provides a closer relationship between the fares and the costs of service relating to various markets would also produce more equitable results among the carriers in terms of the ability of individual carriers to achieve a fair return on investment. Further, fares based upon reasonable costs and load factors would benefit the communities served by discouraging uneconomic scheduling in some markets and encouraging a better balance of service on all the carriers' routes. The establishment of fares on an economic basis would also provide assistance in future determinations of the need for and adequacy of air service to various communities.[36]

Although it would be inaccurate to claim that the Board overwhelmingly endorsed efficient pricing for determining fare structure, on most of the major DPFI issues thus far decided it at least moved in this direction, sometimes breaking precedent with the past.

In the joint fares phase of the DPFI, the major questions were: (1) what discount for through service should be allowed (or required), and (2) in the event there is a discount, what constitutes a reasonable division (among carriers) of the revenue collected from the traveller?[37] In its decision, the Board concluded that on average the carriers save $4.00 per interline connection. (This saving accrues mainly from single ticketing and through-routing of baggage.) Accordingly, the Board ordered carriers to provide a minimum $4.00 discount for each intercarrier connection. The Board also allowed larger discounts for "competitive" interline fares. (These are instances where two carriers, not certificated for direct service between two points, may match the direct-service fare of a third carrier.)

---

[34] The airlines appear to be characterized by constant returns to scale in most output dimensions, and thus, for practical purposes, we may assume that marginal cost equals average cost. For a discussion and some recent evidence, see Douglas and Miller [2], Chapter 2.

[35] Third-degree price discrimination may be (second-best) efficient for a regulated industry facing a cost-recovery revenue restraint, but only if returns to scale are not constant. We wish to distinguish here between price *discrimination* and price *differentiation* (i.e., the latter referring to an efficient structure of peak-load pricing and price/quality options).

[36] CAB Order 70-1-147, p. 5.

[37] Joint fares are rates charged to passengers travelling over the lines of two or more carriers with a single ticket. An example would be a passenger flying from Los Angeles to Charlottesville, Virginia, via Washington, D. C. He might fly from Los Angeles to Washington on TWA, then transfer onto Piedmont for the Washington-Charlottesville portion of the journey. At Los Angeles, he would buy a single ticket which might or might not include a "discount" over the sum of the normal Los Angeles-Washington and Washington-Charlottesville tariffs.

On the question of dividing joint-fare revenue among the participating carriers, the Board chose in principle to price according to the relative costs of providing the service:

> It is an elementary principle of regulation that revenues be geared as closely as possible to the costs of service. With costs as the touchstone of regulation, efficiency is promoted, and inefficient misallocations of resources are discouraged; and as long as rewards are commensurate with reasonably attainable costs, sound economic conditions will be fostered. . . . Since the carriers' inputs of services in interline carriage are represented by the cost of service, a division formula which rewards each carrier in proportion to the cost of the services it provides . . . must be deemed to be a just, reasonable and equitable basis for divisions.[38]

In applying this principle, the Board adopted a line-haul cost formula developed by its Bureau of Economics as the standard for revenue division. Although certainly constituting a rough-and-ready approach, the essence of the joint-fare decision can be viewed as a proxy for efficient pricing.

The discount fare question was an extremely sensitive one for the Board inasmuch as special favors had been meted out to young people (among others), and to withdraw this privilege would give the public appearance of being anti-youth. Furthermore, the Board would be reversing its previous endorsement of discounts, and this would require explicit justification. In spite of these difficulties, the Board did conclude such discounts to be unlawful. Reflecting the DOT arguments, the Board found in essence that discounts are instances of (third degree) price discrimination, that therefore they have allocative efficiency costs, and that concomitant with the abolition of discount fares would be a downward adjustment in the level of normal fares.

Portions of the discount fare decision are worth quoting:

> Assuming that average seat-mile costs remain constant, the industry's breakeven load factor levels will necessarily increase as average fares are reduced and will decrease if average fares increase. . . . The introduction of discount fares reduces the average fare, thus increasing the breakeven load factor. As the Department of Transportation has shown, if the discount fares have the effect of bringing the breakeven load factor above the actual load factor, then the carrier has two choices: either it must adjust its schedules in order to raise the actual load factor to breakeven or the fare level must be raised in order to lower the breakeven load factor. In either event, the normal-fare passenger is burdened. In the first instance, his fare remains the same but there will be fewer seats available to the normal-fare-paying passenger, and there will thus be a decline in the quality and adequacy of the services provided to the normal-fare payer, unaccompanied by any reduction in his fare. On the other hand, if the carrier increases its fares to cover the declining average yield, then the normal-fare payer is burdened by having to bear a fare increase in order to cover the short-fall created by the discount fares.[39]

---

[38] CAB Order 72-4-42 (April 10, 1972), p. 31.

[39] CAB Order 72-12-18 (December 5, 1972), pp. 45 and 46.

The Board went on to say that promotional fares of a short-run duration (essentially those of a "loss-leader," information-generating variety) would be permitted so long as they served their purpose and did not remain in effect for an extended period.[40]

To implement its findings, the Board declared that within an 18-month period all youth, family, and Discover America discount fares would have to be eliminated. Moreover, it recognized that it would be necessary to revise (downward) the level of normal fares and announced its intention to institute another proceeding which would determine the required readjustment.[41]

The remaining fare structure issues were drawn into focus in the "fare structure" phase of the investigation. To this date, the Board has not rendered its final opinion, although we have the Administrative Law Judge's recommended decision. Thus, here we can only describe the issues without knowing their ultimate resolution.

The most important issue of this phase concerned the appropriate "fare taper." As would be predicted by the theory, there is evidence that carriers are breaking even at various lengths of haul, given the present fare structure;[42] moreover, a change in fare structure within reasonable limits would, over time, result again in a zero-profit load factor equilibrium for the various distances. Given that in the short run costs are not very variable and that demand is fairly inelastic, generally it would be to the advantage of a carrier to have its fares raised or for the fare structure to be altered so that fares rise for its weighted-average stage length. Such a fare increase would thus bring about excess profits until which time nonprice competition eliminated them.

As a whole, the carriers accepted as an article of faith that some fare (-mileage) formula would be adopted in the proceeding.[43] Eight different fare formulae were advanced, with the remaining parties indicating a preference for one of these eight. Consistent with the point raised above, it is interesting to note that TWA, which has the longest average length of haul, proposed to *reduce* fare taper and raise fares noticeably in long-haul markets. Braniff, which has the shortest stage length, proposed to increase short-haul fares and decrease long-haul fares.

---

[40] It is worth noting that in the discount fare phase the Board drew tentative conclusions regarding scale economies of particular relevance to the question of airline mergers:

> The effects of scale on the economics of airline operations was considered by the Department of Transportation. That party submitted evidence in support of the proposition that the airline industry is one of constant returns to scale. . . . No party has seriously challenged the DOT evidence, and without necessarily adopting its methodology and conclusions in all of its details, the Board finds that the general conclusions reached are sound. (*Ibid.*, p. 48.)

[41] It might be noted that immediately after the Board's decision on discount fares a number of bills were introduced in Congress which would either grant the Board explicit authorization to approve discounts for youth, elderly, etc, or actually force the Board to have such fares reinstituted. To date, no formal action has been taken on such bills.

[42] With the cost taper (for a constant load factor) exceeding the fare taper, one would predict generally declining load factors with increasing distance. Such is the case—see Douglas and Miller [2], Chapter 5.

[43] Essentially a fixed charge plus a constant (or variable) rate per mile. (Cf., Figure 4.)

Additional issues in this phase of the investigation were brought into focus. For example, should the fare taper incorporate the load factor standards which DOT recommended as consistent with its estimated optimal level and structure of fares? Also, as DOT recommended, should peak-load pricing be encouraged to a greater degree? Should airports be encouraged to ration off landing slots and gates through the price mechanism? What are the appropriate relationships among fares for first-class, coach, and economy services? These issues, as well as the specific fare taper, have yet to be answered by the Board. The Administrative Law Judge's recommended decision is a conservative one, and does not reflect the "enlightenment" of the Board in previous DPFI decisions. A final decision by the Board in this phase of the investigation has, for some time now, appeared imminent.

5. Measures to increase price competition

■ One theme of our presentation was that even in "competitive" markets there is little quality differentiation, and practically no variation in price/quality combinations (except for first-class vs. coach service). If there were price competition, then market tests would lead to a more efficient level and structure of price and quality and also to more price/quality differentiation.

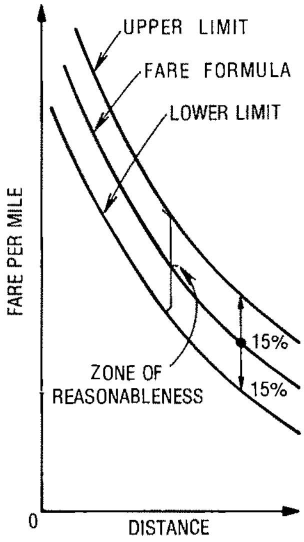
In the fare structure phase of the DPFI, proposals were advanced which would promote some measure of price competition. Essentially, the idea was that the Board should establish a "zone of reasonableness" within which carrier fare filings would be *per se* or else *prima facie reasonable.*[44] This zone of fare flexibility would not be without the usual restraints on discrimination, preference, and prejudice; however, under the proposal, carriers could initiate limited fare changes without fear that they would be suspended and put down for investigation on grounds that the new fares were unreasonably high or unreasonably low.[45]

The zone of reasonableness concept advanced in the case is summarized in Figure 4. Starting with a fare (-mileage) formula, carriers would be allowed to establish and/or change specific fares anywhere within this zone. Fares could also be proposed outside the zone, the only difference being that such fares would not be assumed reasonable and would have to go through the same procedure as at present.

This proposal for pricing flexibility had its most ardent advocates in two government agencies—the Departments of Transportation and Justice. These two executive-branch representatives saw the proposal primarily as a means of encouraging carriers to *price compete* as an alternative to competing on the basis of service quality. They also took the more extreme position on the zone, such that rates within the boundaries would be *per se* (vis-à-vis *prima facie*) reasonable. Under their proposal, rates would be

FIGURE 4

ZONE OF REASONABLENESS PROPOSAL TO INCREASE PRICE COMPETITION



---

[44] Section 1002(d) of the Federal Aviation Act gives the Board the power to ". . . determine and prescribe the lawful rate, fare, or charge (or the maximum or minimum, or *the maximum and minimum thereof*). . . ." (Emphasis added.)

[45] A further restraint, of course, would be Section 411 ([Unfair] Methods of Competition).

allowed to vary ± 15 percent from a predetermined fare (-mileage) formula, thus yielding an established zone "30 percent wide."[46]

Other parties to the investigation also advocated some fare flexibility, but they suggested a zone in which fares would be only *prima facie* reasonable (i.e., presumed reasonable unless proved otherwise). The most outstanding proponent of this scheme was American Airlines, traditionally the industry's pricing leader, who, in proposing a ± 15 percent zone, stated that in regulating fares,

> [t]he Board should . . . attempt to achieve two goals. First, it should give the carriers a real opportunity to develop prices which reflect differences in costs and marketing considerations among different markets. Second, it should give the public the benefit of meaningful price competition.[47]

The other parties advocating pricing flexibility were more conservative in their approach. Continental, Delta, four local service carriers, and the Board's Bureau of Economics all advocated a zone of ± 5 percent.[48] Rather than emphasizing price competition, however, their arguments were predicated primarily on the need for (limited) carrier management discretion in establishing rates.

Of the remaining major parties, only Allegheny and Ozark endorsed the concept of fare flexibility. The rest either ignored it or opposed it outright. Strongest among those in opposition were Eastern, TWA, and United. For example, Eastern argued that,

> The wide disparity among carriers as to inherent strength . . . requires that fares be controlled and maintained at a level which will permit adequate returns to the "need" carriers and protect them from the adverse effect of fare competition with carriers who are stronger . . . .[49]

In essence, those opposed to pricing flexibility voiced fears of "destructive competition" and raised the specter of open "rate wars."

The Administrative Law Judge's recommended decision in the fare structure phase forcefully argued against allowing fare flexibility, saying that by doing so the Board would abdicate its fare-regulating responsibilities. The Board has yet to rule on this issue. However, there appeared to be some preliminary indication that it might find some degree of fare flexibility acceptable. In its fare level decision, the Board agreed with American that in certain short-haul markets it might be expedient to retain the existing fare instead of raising it the authorized 9 percent.[50] In effect, however, this decision really authorized the carriers to select either the old fare or to take the full 9-percent increase. The "zone" is narrow (only 9 percent, or ± 4.5 percent) and is limited to the extremes. Since a significant fare increase had been wholly anticipated and capitalized, the carriers immediately raised nearly all such rates the authorized amount.[51]

---

[46] Obviously, this was a means of partially deregulating the airlines. In commenting on the earlier version of this paper [3], Richard Caves remarked, "Denied the option to deport the CAB wholesale, Miller and [Douglas] have chosen instead to offer it a lifetime supply of sleeping pills."

[47] See Brief of American Airlines, Inc. to the Civil Aeronautics Board, CAB Docket 21866-9 (May 22, 1972), p. 16.

[48] The Moss Group (of Congressmen) advocated a zone of minus 15 percent.

[49] See Brief of Eastern Air Lines, Inc. to the Civil Aeronautics Board, CAB Docket 21866-9 (May 22, 1972), p. 42.

[50] "American Airlines has proposed that any authorized fare increase be expressed as a maximum fare level. Under a maximum fare order, the carriers would be permitted to raise their fares to the prescribed level but

TABLE 1

COMPARISON OF POLICIES GPFI (1960), POLICIES DEVELOPED DURING INTERIM, AND DPFI (1972)

POLICY ISSUE (PHASE OF DPFI)	GPFI (1960)	POLICIES DEVELOPED DURING INTERIM	DPFI (1972)
1 FLIGHT EQUIPMENT INVESTMENT	STRAIGHT-LINE WRITEOFF	—	AFFIRMED
2 LEASED AIRCRAFT	—	NOT INCLUDED IN RATE BASE	AFFIRMED
3 DEFERRED FEDERAL INCOME TAXES	INCLUDED AS CURRENT EXPENSE	—	AFFIRMED
4 JOINT FARES	— —	a RATE BASED ON SUM OF LOCAL RATES MINUS COST SAVING b REVENUE DIVISION BASED ON LOCAL RATES	a RATE BASED ON SUM OF LOCAL RATES MINUS COST SAVING b REVENUE DIVISION BASED ON RELATIVE CARRIER COSTS
5 DISCOUNT FARES	—	DISCRIMINATORY DISCOUNTS ENCOURAGED	DISCRIMINATORY DISCOUNTS DISALLOWED
6A SEATING CONFIGURATION STANDARDS	—	—	STANDARDS ADOPTED, REGULATION IMPLICIT
6B LOAD FACTOR STANDARDS	REJECTED	—	STANDARDS ADOPTED, REGULATION IMPLICIT, NO ANALYSIS OF PRICE VS QUALITY
7 FARE LEVEL	UNABLE TO DETERMINE REASONABLE FARE LEVEL FROM RECORD	FARES BASED ON ACTUAL COSTS, REFLECTING CARRIER "REVENUE NEED"	FARES BASED ON AVERAGE COST OF PROVIDING REASONABLE QUALITY OF SERVICE
8 RATE OF RETURN	a 10 5 PERCENT OVERALL RETURN, BASED ON COST OF CAPITAL AND ACTUAL DEBT/EQUITY RATIO b SOME IMPLICATION OF GUARANTEED RETURN	— —	a 12 PERCENT OVERALL RETURN, BASED ON COST OF CAPITAL AND "OPTIMAL" DEBT/EQUITY RATIO b "GUARANTEE" EXPLICITLY REJECTED, ACTUAL RETURN LEFT TO MARKET FORCES, NO "RATCHET EFFECT"
9 FARE STRUCTURE	—	GRADUAL INCREASE IN FARE TAPER, PROMULGATION OF INDUSTRY-WIDE FARE FORMULA, NO ZONE OF REASONABLENESS FARE FLEXIBILITY	(BOARD DECISION NOT YET RENDERED, SOME PRELIMINARY INDICATION FAVORING COST-BASED RATES AND LIMITED PRICING FLEXIBILITY)

SOURCES CAB ORDERS E-16068, 70-1-147, 71-4-48, 71-4-54, 71-4-58, 71-4-59, 71-4-60, 72-4-42, 72-12-18, 72-5-101, *et al.*

---

would not be required to do so. . . . The Board believes that American's proposal has considerable merit, . . . [and] . . . that there is a sufficient public interest in permitting the carriers a degree of discretion in implementing the authorized fare increase as to warrant a trial of the maximum fare concept." (CAB Orders 71-4-59 and 71-4-60, *ibid.*, pp. 78–80.)

[51] It should be noted that a zone which is too narrow not only limits the degree of price competition but also would facilitate price collusion. Thus, to have the desired effect, the Board's decision must include a broad

6. Summary of DPFI policy changes

■ Table 1 summarizes the findings of the DPFI in comparison with the outcome of the GPFI of 1960. Since questions about fare structure were explicitly set outside the scope of the GPFI (and thus no opinions were rendered at that time on such issues as joint fares, discount fares, and fare taper), additional comparisons are drawn based on policies developed subsequent to the GPFI.

As indicated in the table, on a number of important issues the Board clearly broke with tradition (e.g., Phases 5, 6, 7, and 8). Generally, these changes may be attributed to two underlying policy emphases which have evolved from the investigation. First, there has been an acceptance by the Board of the nonprice competing cartel model outlined earlier. Second, greater emphasis has been placed on costs as the standard for individual fares; accordingly, there is less emphasis on "value of service" and other subjective criteria. We must note, however, that the Board has not and likely will not allow (and encourage) extensive price competition among the carriers, although it may appear to be relaxing these restraints to some extent. Moreover, the DPFI dealt only with fare issues, not with questions of entry and exit. Arguably, a general investigation of entry policy would result in more liberal standards. Thus, while the policy actions taken in the DPFI are encouraging, there remain many additional policy changes which would appear warranted on grounds of economic efficiency.

7. Postscript

■ We note in conclusion two final points. First, while the outcome of the DPFI is encouraging in terms of what it might portend for future CAB regulatory performance, we must stress that the Federal Aviation Act allows the Board considerable latitude in its policies and that regulatory philosophy at any point in time is greatly dependent upon the membership of the Board. Since the major decisions were rendered in that investigation, the Chairmanship of the Board has changed hands, and one position remains vacant. According to recent pronouncements by the new Chairman, the Honorable Robert D. Timm, the Board will henceforth concentrate on maintaining a "healthy industry."[52]

Several actions have been taken. Of potentially far-reaching importance, the Board has recently authorized and even *encouraged* the carriers to reach agreements on controlling schedule capacity and has stated that such agreements would appear to be an appropriate and successful "regulatory device."[53] This attitude is in contrast with the extreme reluctance with which the Board approached the matter when capacity agreements were first proposed. Said the Board's initial order:

---

range for price action; we expressed the opinion that ± 15 percent represents a minimal standard.

[52] Cf., Robert Timm's Luncheon Address to the Aero Club of Washington, D. C. (April 24, 1973), "An Industry Champion Pilots the CAB," *Business Week* (May 26, 1973), pp. 76–77; and Albert R. Karr, "Under New Chief, CAB Bids to Lift Profits of Airlines, Annoying Consumer Advocates," *Wall Street Journal* (August 7, 1973), p. 32.

[53] Cf., CAB Order 73-4-98 (April 24, 1973).

> We regretfully have concluded in view of the evidence of low load factors in the major markets in question that [economic forces] are acting too slowly in the immediate circumstances. In our judgment, a limited departure from our normal policy of leaving scheduling and capacity to the free play of competitive forces is justified at this time. . .[54]

Succeeding Board orders, however, have backed away from this stance, culminating in a recent decision to,

> put all parties on notice that the Board has tentatively concluded that capacity limitation agreements may in many circumstances be in the public interest and that the Board is prepared to consider requests for authority to discuss capacity reduction agreements in specific markets and to judge resulting agreements on their merits.[55]

At present a whole range of issues surrounding capacity agreements is under investigation.[56] However, there is little doubt that a presumption exists at the Board (especially on the part of its Chairman) in favor of their utilization.

In other related actions, the Board summarily reversed its Phase-5 DPFI decision to reduce the level of normal fares, and instead merely ordered the phase-out of discount fares over an 18-month period.[57] On alleged energy conservation grounds, the Board authorized the carriers to discuss agreements leading to fuel conservation through reductions in aircraft speed.[58] In order to limit perceived "destructive competition" and "substantial impairment" of air services, the Board has permitted the U. S. international scheduled and supplemental carriers to negotiate an agreement setting minimum prices for trans-Atlantic charter flights.[59] Among specific measures envisioned are: (1) a slow-down of new route authorizations, (2) decertification of carriers in "over-competitive" markets, (3) a furthering of the policy of cross-subsidizing small city-pair markets, and (4) across-the-board fare increases.[60] Of course, these measures not only mark a retreat from the "enlightened" policy direction which we have imputed to the DPFI, but they are also anticompetitive and have associated significant efficiency costs. Whether such policies will in fact be continued and/or promulgated remains to be seen.[61]

The second point we wish to note is that the analysis developed for this case should be applicable to a broad variety of other industries and regulatory phenomena. Obvious candidates are other

---

[54] CAB Order 71-8-91 (August 19, 1971), p. 5. Also, see *ibid.*, pp. 3–10, especially p. 3.

[55] CAB Order 73-4-98, p. 1.

[56] See CAB Order 73-7-147 (July 27, 1973) which approved the latest three-carrier transcontinental capacity agreement for an interim period of six months and set down an investigation of the capacity limitation concept.

[57] See CAB Order 73-5-2 (May 1, 1973).

[58] See CAB Order 73-5-123 (May 25, 1973). Said the Board, "[A]greed-on flight times would lead to carrier observance of conservation procedures because no competitive advantage would be gained by beating such flight times." (*Ibid.*, p. 1.)

[59] See CAB Order 73-6-79 (June 19, 1973).

[60] See note 52.

[61] In November, 1973, despite (Board-orchestrated) schedule cut-backs which lowered costs but only marginally affected revenues, the Board approved an across-the-board fare increase of 5 percent. [See CAB Order 73-11-93 (November 21, 1973).]

"competitive" industries under regulation. These include, for example, the motor trucking industry, the water carrier industry, commercial banking, state regulation of insurance, and the regulation of fees paid to securities and commodity brokers. Each has the characteristic of (approximately) zero industry profits at a range of feasible price/quality combinations. Likewise, other semiregulated, or cartelized industries may be described with this analysis. These include physicians, real estate agents, attorneys, hotel accommodations, and service industries where prices are determined by "union scale" (e.g., barber shops). In each case there would appear to be price collusion, but control of output is imperfect; welfare losses would be indicated not only by excess profits, but by nonoptimal price/quality options. Finally, oligopolistic industries where prices are sticky likewise may be brought into question regarding price/quality options, and this is a dimension of efficiency analysis not yet developed in the industrial organization literature.

[On March 18, 1974, the Board issued its decision in the fare structure phase of the DPFI (CAB Order 73-3-82). In it the Board rejected the zone of reasonableness concept, stating that ". . . adoption of any of these proposals would be tantamount to abdicating our statutory responsibility to protect the public interest . . ." (p. 115). However, the Board did disavow the notion of cross-subsidy and enunciated a policy of basing the fare taper on the cost taper: ". . . the Board's long-run goal must therefore be oriented towards a fare structure based on the industry average costs of service at the various lengths of haul . . ." (p. 68). On a related issue, the Board neither accepted nor rejected the Douglas-Miller proposal regarding the optimal level and distance structure of load factors: "DOT's analysis which . . . attempts to define a reasonable trade-off between fares and quality of service from the *passengers'* viewpoint, has considerable theoretical merit. However, its application depends on a determination of the value passengers place on their time, and the record is not sufficient to support any findings on this crucial element of DOT's analysis" (emphasis in original; p. 42).]

References

1. Douglas, G. W. "Testimony." CAB Docket 21866-9, DOT-T-3, May 17, 1971.
2. ——— and Miller, J. C., III. *Economic Regulation of Domestic Air Transport: Theory and Policy.* [A portion of this study will appear as, "Quality, Competition, Industry Equilibrium, and Efficiency in the Price-Constrained Airline Market." *The American Economic Review*, Vol. 64, No. 1 (March 1974).] Report prepared for the Brookings Institution, processed May 1973.
3. Miller, J. C., III. "The Domestic Passenger Fare Investigation and the Future of CAB Regulation." Paper presented at an AT&T Seminar on Problems of Regulation and Public Utilities, Amos Tuck School of Business Administration, Dartmouth College, August 30, 1971.
4. Redford, E. S. "An Administrative Center at Work: The CAB's General Passenger Fare Investigation," in Edwin A. Bock, ed., *Government Regulation of Business: A Casebook*, New York, Prentice-Hall, Inc.: 1965.

Peak-load pricing: some further remarks

Oliver E. Williamson

Professor of Economics
University of Pennsylvania

This note takes issue with the "reconciliation" proposed by George Wilson between the Steiner and Williamson formulations of the peak-load pricing problem. Wilson respecifies weights without simultaneously addressing himself to the question of expressing output units in consonant terms. This procedure is incorrect, whence the error associated with the original Steiner formulation cannot be overcome in this way. Steiner's interpretation of his results in price discrimination rather than marginal cost pricing terms is briefly discussed. Reasons for preferring the marginal cost pricing interpretation are offered.

1. Introduction

■ Two comments on the peak-load pricing problem have appeared recently which take issue with my earlier treatment of this question. Since both seem to me to be contestable in certain respects, and as others have inquired over my response to these remarks, it may be useful to add still another comment to this literature. I emphasize, however, that my remarks are strictly pedagogical; they do not advance the analysis of peak-load pricing, but rather are intended to tidy up some loose ends. The interpretations which concern me are those of George Wilson[1] and Peter Steiner.[2]

2. Specifications

■ Wilson's comment attempts a "formal reconciliation" between my model[3] of the peak-load pricing problem and Steiner's earlier contribution.[4] One of the devices that he uses for this purpose is to interpret the weights which appear in my model in different terms. Thus whereas the w_i which appear in my model refer to the proportion of the entire demand cycle over which the demand of type i prevails, with $\sum_{i=1}^{n} w_i = 1$, Wilson indicates that my results become identical with Steiner's if $w_i = 1$, all i.

It is true that for the special case where (1) all periodic loads are of equal length and (2) only one load utilizes plant to capacity that my peak-load pricing results can be made *to appear*

Research on this paper has been supported by a grant from the National Science Foundation.

[1] In [7].

[2] In [5].

[3] See [6].

[4] See [3].

identical to Steiner's in this way. But there is more to it than respecifying weights. My original weights hold for the case where the abscissa reads in terms of units per cycle. Under Wilson's respecification, the abscissa must be expressed as units per (equal length) subperiod. But Wilson indicates no awareness of the need to express the weights used and the units of the abscissa *in consonant* terms. He respecifies the one without making reference to the concurrent need to respecify the other.[5]

Wilson goes on to indicate that the Q_i in my expressions represent "sales per subunit of a subperiod of the total cycle" and that these are "sales for a unit of time that is less than the period of the peak or offpeak."[6] This is simply incorrect. My demand curves are expressly defined in terms of the entire cycle. Thus, "demand is expressed as $D_i^{(w_i)}$, where i refers to the subperiod in question and superscript (w_i) to the fraction of the cycle during which the demand in question prevails, with each demand curve showing the amount of output per cycle which would be demanded at every price were the demand in question to prevail over the entire cycle."[7] Accordingly, the abscissa in my figures is expressed in output per cycle terms.

This is not to say, however, that the quantity units could not be defined differently. Thus suppose that one hour is the shortest interval for which any periodic load will prevail and that the length of a cycle is twenty-four hours. Then demand could be expressed as $D_i^{(h_i)}$, where h_i is the number of hours for which demand of type i prevails and $h_i = 24w_i$. Let the abscissa here be denoted by q while that in my original diagrams and related expressions is given by Q. Then $Q/24 = q$ and delivered output at price P_i during period i is $Q_iw_i = q_ih_i$. The n-period partial equilibrium welfare maximization problem[8] can thus be conveniently formulated in either of two equivalent ways:

---

[5] If, in the models specified in (1) and (1′), below, q_i is constrained to take on the same numerical value as Q_i under both weighting schemes, then a different symbol must be used to designate short-run marginal cost (assuming $b \neq 0$) for the two weighting schemes and different unit prices are implied. Inasmuch as Wilson denotes SRMC by b per unit throughout, the above indicated constraint does not apply and the relation $q_ih_i = Q_iw_i$ holds instead.

[6] [7], p. 309.

[7] [6], p. 817.

[8] To illustrate, consider the two-period case for which demand of type 1 prevails for one-third of the cycle and demand of type 2 prevails over the remaining two-thirds. Then $w_1 = 1/3$, $w_2 = 2/3$, and $h_1 = 1$, $h_2 = 2$. Assume that these demands are independent and that both utilize plant to capacity. The expressions for optimal capacity are then obtained by setting $Q_1 = Q_2 = Q^*$ and $q_1 = q_2 = q^*$, differentiating expressions (1) and (1′) in the text by Q^* and q^*, respectively, and setting equal to zero. The following expressions result:

$$P_1(1/3) + P_2(2/3) = b + \beta \tag{i}$$

$$P_1(1) + P_2(2) = b(3) + \beta(3). \tag{i'}$$

Suppose, in particular, that demands are given as $D_1^{(1/3)} = 30 - 1/2Q_1$ and $D_2^{(2/3)} = 60 - Q_2$ or, equivalently, that $D_1^{(1)} = 30 - 3/2q_1$ and $D_2^{(2)} = 60 - 3q_2$. Suppose, furthermore, that $b = 10$ and $\beta = 30$. Then the following price and output relations obtain:

$$\max W = \sum_{i=1}^{n} (TR_i + S_i) w_i - b \sum_{i=1}^{n} Q_i w_i - \beta Q^*, \quad \text{or} \tag{1}$$

$$\max W = \sum_{i=1}^{n} (TR_i + S_i) h_i - b \sum_{i=1}^{n} q_i h_i - \beta q^* \sum_{i=1}^{n} h_i, \tag{1'}$$

$$\text{where} \sum_{i=1}^{n} w_i = 1,\ q_i h_i = Q_i w_i, \quad \text{and} \quad q^* \sum_{i=1}^{n} h_i = Q^*.$$

There is still a third way to handle the unit specification problem. This is to have the demand curves "unweighted," in any explicit sense, but to redefine the units of the abscissa whenever a change is made in the number of (equal length) periodic loads. Demand in period i is simply denoted by D_i, and the units of the abscissa are appropriately expressed as a function of the total number of periods, n.

This third interpretation, which did not occur to me initially (since I thought it natural to work with standard units throughout), is perhaps closest to Steiner's formulation. The problem with Steiner is that he makes no mention of the need to respecify the abscissa in this way.[9] In relation, however, to the Steiner paper taken as a whole, the resulting error is surely negligible—and I did not mean to exaggerate it then nor do I now. That, fifteen years later, the Steiner paper still constitutes for many students

---

$$P_1 = 24,\ P_2 = 48,\ Q^* = 12 \tag{ii}$$

$$P_1 = 24,\ P_2 = 48,\ q^* = 4. \tag{ii'}$$

The corresponding revenues are:

$$P_1 Q^* w_1 = 96,\ P_2 Q^* w_2 = 384, \tag{iii}$$

$$P_1 q^* h_1 = 96,\ P_2 q^* h_2 = 384, \tag{iii'}$$

while costs are:

$$(b + \beta) Q^* = 480, \tag{iv}$$

$$(b + \beta) q^* (h_1 + h_2) = 480. \tag{iv'}$$

Given the constant returns to scale technology involved, breakeven naturally obtains.

[9] Thus consider Steiner's Figures 1 and 2 in [3]. D_1 in both diagrams shows the amount of output that will be demanded at every price during the first period. Output in both diagrams is given by a common symbol X while price is denoted by P. Absent any indication to the contrary, for which there is none, one presumes that the units in both diagrams are identical. But this will not do; demands apply for half a cycle in the two-period (Figure 1) case, but for a third of a cycle in the three-period (Figure 2) condition. Unless the demand curves themselves are weighted, the units on the abscissa must be redefined to reflect these changes in load duration. (This assumes that the duration of the cycle is fixed, which is the usual assumption.) Nowhere in the original Steiner paper is the problem of compatible units acknowledged, much less addressed; whence the ambiguity in his formulation of the problem that I pointed out in my original article—Wilson's comments to the contrary notwithstanding. (I did not, however, initially attribute the difficulty to the specification of output units. I assumed that these were to be held constant and interpreted the difficulty in cost specification terms.)

an insightful introduction to the peak-load pricing issue stands, I trust, as testimony to its basic merits.

3. "Price discrimination"

■ Steiner interpreted his peak-load pricing results in price discrimination terms. Jack Hirshleifer took issue with this and suggested instead that they could be more usefully interpreted in marginal cost pricing terms.[10] Steiner objected,[11] apparently still does,[12] and observes that my expressed sympathy with the Hirshleifer position is gratuitous.[13]

Perhaps I can explain. My inclination, when I was working on the peak-load pricing paper in 1965, was to ignore this dispute. I had the benefit, however, of Jack Hirshleifer's comments on a late draft of that paper, and he suggested that I indicate where I came out on the issue. It seemed to me obvious from the analysis in the paper that I favored the Hirshleifer view. I nevertheless added a footnote in the last version to that effect.

It will be recalled that Steiner stated early in his original article that previous authors had been "led to the incorrect conclusion that an optimal pricing scheme necessarily lies, not in some scheme of discriminatory pricing, but in a sophisticated application of marginal cost pricing."[14] Steiner sought to correct this error by interpreting his pricing scheme in price discrimination terms. Hirshleifer objected that Steiner's definition of price discrimination was unconventional and argued that the marginal cost pricing analog should be maintained. I interpreted my results in Hirshleifer's terms.

One way of putting the issue is as follows: Is the uniform load pricing problem more usefully regarded as a special case of sequential load discriminatory pricing or as a special case of sequential load marginal cost pricing? Since price discrimination in conventional terms turns on demand elasticity considerations, while in Steiner's terms it varies with the level of demand but not the elasticity, the Steiner analog seems to me misleading. By contrast, uniform load marginal cost pricing can be shown to be the special case of sequential load pricing for which $w_k = 1$ and $w_i = 0$, all $i \neq k$.[15] Reference to price discrimination tends to disguise the relation between uniform load marginal cost pricing and optimal (partial equilibrium) variable load pricing. Inasmuch as students are often put off by discussions of partial equilibrium marginal cost pricing when it is formulated as a uniform load problem, but find the rationale more compelling when it is developed in a sequential load context, choice of terms seems to be of some pedagogical significance.

---

[10] In [2].

[11] See [4].

[12] See [5], pp. 19–20.

[13] [5], p. 19.

[14] [3], p. 586.

[15] More generally, the uniform load condition is given by $0 < w_k < 1$, $w_i = 0$, all $i \neq k$. That the uniform load marginal cost pricing results on page 812 of [6] are a special case of the sequential load condition can be seen by substituting $w_2 = 1$, all $w_i = 0$, $i \neq 2$, into the formulae on pp. 820–821 of [6].

The issue can be posed in a second and somewhat more rigorous way by considering two alternative specifications of the objective function—namely, maximize profits or, alternatively, maximize the conventional partial equilibrium welfare function[16]—and by comparing the pricing relations that obtain when *simultaneous* (yet distinguishable) demands prevail with those that obtain for the *sequential* demand problem. The simultaneous demand condition is the familiar one that is treated in the textbooks; the pricing results derived for this case have an unambiguous interpretation. The sequential demand problem is the peak-load problem; the pricing relations derived here are the ones that are in dispute.

Consider first the simultaneous demand situation. The alternative objective functions for this case are given by:

$$\max \pi = P_1Q_1 + P_2Q_2 - C\,(Q_1 + Q_2) \tag{2}$$

$$\max W = (TR_1 + S_1) + (TR_2 + S_2) - C\,(Q_1 + Q_2), \tag{2'}$$

where $C(Q_1 + Q_2)$ expresses cost as a function of total output. Assume, for simplicity, that $C(Q_1 + Q_2) = (b + \beta)(Q_1 + Q_2)$, and set the partial derivatives of π and W with respect Q_1 and Q_2 equal to zero. Noting that $\frac{\partial P_i}{\partial Q_i} = -\frac{P_i}{Q_i} \cdot \frac{1}{\eta_i}$, where η_i denotes price elasticity, the following pricing expressions are obtained:

$$\begin{cases} P_1 = \dfrac{b + \beta}{1 - 1/\eta_1} \\ P_2 = \dfrac{b + \beta}{1 - 1/\eta_2} \end{cases} \tag{3}$$

$$\begin{cases} P_1 = b + \beta \\ P_2 = b + \beta. \end{cases} \tag{4}$$

The prices shown in (3), which correspond to the profit maximization formulation, involve familiar third-degree price discrimination. The prices shown in (4), which obtain when welfare is maximized, reveal that prices are here set equal to marginal costs for both demand classes.

Suppose now, instead of simultaneous demands, that we have sequential demands to contend with. Assume also, for simplicity, that plant is used to capacity only in the second period. The corresponding profit maximization and welfare maximization expressions to (2) and (2′) then are given by:

$$\max \pi = P_1Q_1w_1 + P_2Q_2w_2 - bQ_1w_1 - bQ_2w_2 - \beta Q_2 \tag{5}$$

$$\max W = (TR_1 + S_1)w_1 + (TR_2 + S_2)w_2 - bQ_1w_1 - bQ_2w_2 - \beta Q_2. \tag{5'}$$

Taking derivatives, as above, and solving for the prices that obtain in each case, we get:

---

[16] My formulation of the problem in this way was stimulated by my reading of a draft version of the paper by Elizabeth Bailey and Lawrence White [1].

$$\begin{cases} P_1 = \dfrac{b}{1 - 1/\eta_1} \\ P_2 = \dfrac{b + \beta/w_2}{1 - 1/\eta_2} \end{cases} \tag{6}$$

$$\begin{cases} P_1 = b \\ P_2 = b + \beta/w_2, \end{cases} \tag{7}$$

where the prices shown in (6) and (7) correspond to the profit maximization and the welfare maximization formulations of the problem, respectively.

There is no disagreement between Steiner on the one hand and Hirshleifer and myself on the other that the problem which Steiner formulated is that shown in (5′), the solution to which is given by (7). The question is whether it is useful to regard these results as involving price discrimination. I submit that it is not. Rather, the prices shown in (7) very much have the appearance of those obtained in (4) and are to be regarded as marginal cost prices for the sequential load condition. The prices shown in (6), by contrast, do involve price discrimination, in that prices here are adjusted to reflect demand elasticities—as they are in (3). Put differently, the notion of price discrimination among both simultaneous and sequential demand classes is usefully invoked in interpreting equations (3) and (6). But, as everyone agrees, it does not apply at all to the simultaneous demand prices in (4), and, as Hirshleifer and I contend, it does not apply to the sequential demand prices in (7) either.

References

1. Bailey, E. E. and White, L. J. "Reversals in Peak and Offpeak Prices." *The Bell Journal of Economics and Management Science,* Vol. 5, No. 1 (Spring 1974), pp. 75–92.
2. Hirshleifer, J. "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics,* Vol. 72, No. 288 (August 1958), pp. 451–462.
3. Steiner, P. O. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics,* Vol. 71, No. 285 (August 1957), pp. 585–610.
4. ———. "Peak Loads and Efficient Pricing: Reply." *Quarterly Journal of Economics,* Vol. 72. No. 288 (August 1958), pp. 465–468.
5. ———. "Peak-Load Pricing Revisited," in H. M. Trebing, ed., *Essays in Public Utility Pricing and Regulation,* East Lansing, Mich.: Michigan State University, 1971.
6. Williamson, O. E. "Peak Load Pricing and Optimal Capacity under Indivisibility Constraints." *The American Economic Review,* Vol. 56, No. 4 (September 1966), pp. 816–827.
7. Wilson, G. W. "The Theory of Peak-Load Pricing: A Final Note." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 307–310.

Depreciation in a simple regulatory model: comment

Peter B. Linhart

Treasury Department
American Telephone and Telegraph Company

In a recent paper, Jaffee concludes that a regulated firm, seeking to maximize the present worth of cash flows, should adopt a certain form of decelerated depreciation. He assumes the firm is financed solely by debt, and does not take account of the constraint that earnings must at least cover interest. The present note considers a firm financed by a mixture of debt and equity, and includes the interest constraint; it is shown that under these circumstances the optimal depreciation schedule is fastest during the earlier portion of the life of an item of plant.

■ The conclusion of Bruce Jaffee's paper "Depreciation in a Simple Regulatory Model"[1] seems surprising. He concludes that the firm whose objective is to maximize discounted cash flow should choose a depreciation schedule which increases over the life of a "single 'lump' of capital equipment"—i.e., *decelerated* depreciation. Since the problem is an interesting one, it seems worthwhile to see what assumptions led to this conclusion.

First we note that it is clear from Jaffee's equations that the item of plant was financed entirely by debt. Since his definition of cash flow is *net of interest*, the quantity whose present worth is being maximized is just book depreciation.

Second, the firm should have at least enough earnings in each period to pay interest on its debt, although Jaffee does not use this constraint. If we take account of it, we get (in his notation), the inequalities

$$crx_1 \leqslant \text{earnings}_t \leqslant s\left(cx_1 - \sum_{\tau=0}^{t} d_\tau\right), \tag{1}$$

where

x_1 = Amount of capital equipment,
c = Purchase price per unit of this equipment,
r = Interest rate
d_τ = Depreciation in period τ, and
s = Maximum allowed rate of return.

The earnings in period t are given by

$$\text{earnings}_t = R(x_1) - d_t - T_0[R(x_1) - crx_1 - d_t]. \tag{2}$$

The author wishes to thank J. M. Bank, H. A. Eakin, H. J. Landau, and C. O. Lee for helpful discussions.

[1] In [1].

Here R is revenue, assumed the same in each period, and T_0 is the tax rate.

Note that the inequalities (1) cannot possibly be satisfied, since the left-hand member is constant over the life of the item (from 0 to T), while the right-hand member is diminishing, and will in fact diminish to zero.

This contradiction arises from the fact that Jaffee implicitly makes two contradictory assumptions: the assumption of constant interest implies that the debt capital is paid back to the investor only at the end of the life, while the regulatory constraint is written in such a way as to imply that capital is paid back according to the depreciation schedule. Thus, the conclusion of decelerated (sinking fund) depreciation is in doubt.

It is not clear why the objective of the firm should be to maximize the present worth of cash flow, defined in this particular way. Let us accept this objective, however, and ask how the model may be extended to avoid the difficulty outlined above. This may be done in a number of ways; we pick one, as follows.

Suppose the item was purchased with a mixture of debt and equity, the ratio of debt to total capital being δ. Suppose the debt is paid back at the end of the life, but the equity is paid back proportionally to the depreciation schedule. Suppose, further, that the regulatory constraint applies to the return on equity. Let us assume, finally, an arbitrary discount rate, σ, to be used in computing present worth. We now formulate this new problem mathematically.

By a suitable choice of units, we may set the cost parameter, c, equal to unity. We shall use continuous-time notation, just for ease of exposition. Our objective is now to choose a depreciation function, renamed $D(t)$, so as to

$$\underset{D(t)}{\text{maximize}} \int_0^T e^{-\sigma t}\{R - rx_1\,\delta - T_0[R - rx_1\,\delta - D(t)]\}dt, \quad (3)$$

where R is the (level) revenue. Since everything in (3), except the function $D(t)$, is a given constant, we may restate our objective as:

$$\underset{D(t)}{\text{maximize}} \int_0^T e^{-\sigma t} D(t)\,dt. \quad (4)$$

There are three constraints. First, the total depreciation over the life must equal the initial investment:

$$\int_0^T D(t)\,dt = x_1. \quad (5)$$

Second, it must be possible to pay the interest in every period. Since interest is tax-deductible, this is equivalent to requiring that interest be less than pretax earnings:

$$rx_1\,\delta \leqslant R - D(t). \quad (6)$$

Finally, the rate of return on equity must be less than some allowed rate, which we denote by s_e:

$$R - D(t) - T_0 [R - D(t) - rx_1 \delta] - rx_1 \delta \leqslant s_e (1 - \delta) \int_t^T D(\tau) d\tau. \qquad (7)$$

(Here we have used the fact that the regulatory rate base at any time is the amount of equity still to be depreciated.)

We may rewrite (6) and (7) as follows:

$$D(t) \leqslant A \quad \text{(interest constraint)} \qquad (8)$$

$$D(t) \geqslant A - B \int_t^T D(\tau) d\tau \quad \text{(regulatory constraint)}, \qquad (9)$$

where we have introduced the abbreviations:

$$A = R - rx_1\delta$$

$$B = s_e(1 - \delta)/(1 - T_0).$$

The regulatory constraint may be stated more simply by noting that the solution of (9), *under the equality condition*, is:

$$D(t) = Ae^{B(t-T)},$$

so that (9) is equivalent to

$$D(t) \geqslant Ae^{B(t-T)}. \qquad (10)$$

The conditions (8) and (10) are shown in Figure 1. $D(t)$ must lie within the shaded region.

It is evident that for an arbitrary given R, there may be no solution satisfying the constraints. If R is too large, the regulatory constraint will somewhere be violated; if R is too small, the interest constraint will somewhere be violated. We may write

$$R = \rho_e(1 - \delta)x(t) + rx_1\delta + D(t) + T_0[R - rx_1\delta - D(t)],$$

whence

$$R = \frac{\rho_e(t)(1 - \delta)}{1 - T_0} x(t) + rx_1\delta + D(t). \qquad (11)$$

Here $\rho_e(t)$ is the realized return on equity, and $(1 - \delta)x(t)$ is the undepreciated equity at time t. If we require $\rho_e(t) \geqslant 0$, then integrating (11) from 0 to T yields the lower bound on R:

$$R \geqslant \frac{x_1}{T} + rx_1\delta. \qquad (12)$$

The regulatory constraint, $\rho_e(t) \leqslant s_e$, transforms (11) into

$$R \leqslant Bx(t) + r\delta x_1 + D(t). \qquad (13)$$

Taking the present worth of the inequality (13), at discount rate B, gives the upper bound on R:

$$R \leqslant rx_1\delta + \frac{Bx_1}{1 - e^{-BT}}. \qquad (14)$$

It is not difficult to show that (12) implies that

$$\int_0^T D(\tau) d\tau \geqslant x_1,$$

FIGURE 1

PERMITTED REGION FOR DEPRECIATION SCHEDULE

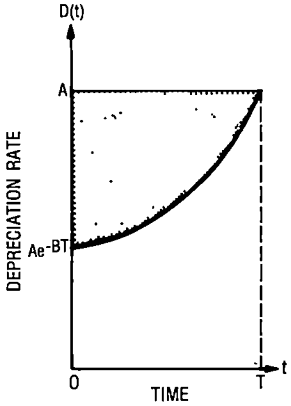
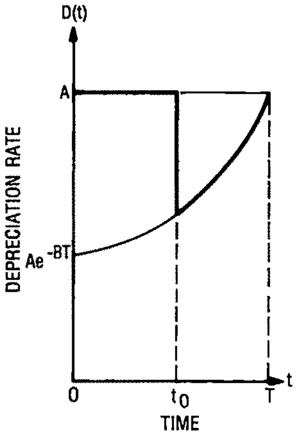


FIGURE 2

OPTIMAL DEPRECIATION SCHEDULE



where the integral is taken along the upper boundary of the permitted region in Figure 1. Similarly, (14) implies that along the lower boundary

$$\int_0^T D(\tau)d\tau \leqslant x_1.$$

Therefore, it is possible to find a depreciation curve which lies in the permitted region and satisfies (5). It is intuitively clear (and can be formally shown) that of such allowed depreciation schedules the one which achieves the objective (3), independently of the assumed discount rate σ, is as shown in Figure 2. Depreciation proceeds at a uniform rate as fast as possible, subject to the interest constraint, until $t = t_0$; thenceforth, it proceeds as slowly as possible, subject to the regulatory constraint. The instant t_0 is determined by (5); we need not quote an explicit formula.

The depreciation schedule shown in Figure 2 can be called accelerated, contrary to Jaffee's result.

Of course, in reality depreciation is supposed to follow the diminution of the economic value of the asset. Nothing is said about this in Jaffee's paper, nor in our comment.

Reference

1. JAFFEE, B. L. "Depreciation in a Simple Regulatory Model." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 338–342.

Depreciation in a simple regulatory model: reply

Bruce L. Jaffee

Assistant Professor
Department of Business Economics and Public Policy
Indiana University

Several difficulties remain with the inclusion of depreciation in a regulatory model, even after Linhart's useful comments and extensions. His liquidity constraint is overly restrictive, there are difficulties in selecting the optimal discount rate(s), and the classification of a depreciation path as uniformly accelerated is incorrect.

■ Peter B. Linhart[1] has developed an interesting extension and presented a number of useful comments on my earlier *Bell Journal* article.[2] However, it is worthwhile to note a few remaining complications.

Linhart is correct in stating that the model should include a liquidity constraint to insure that the firm will not go bankrupt. His equation (6), however, which requires that pretax earnings exceed interest expense in *each* period is more restrictive than necessary. For the completely debt-financed firm, a weaker liquidity constraint would require only that discounted pretax earnings exceed the discounted interest payments. Assuming the firm can borrow and lend at interest rate r, we have

$$\sum_{t=0}^{T} \frac{R(x_1) - d_t}{(1+r)^t} \geqslant \sum_{t=0}^{T} \frac{crx_1}{(1+r)^t}, \quad \text{or}$$

$$[R(x_1) - crx_1]\left(\frac{1+r}{r}\right)\left[1 - \left(\frac{1}{1+r}\right)^T\right] - \sum_{t=0}^{T} \frac{d_t}{(1+r)^t} \geqslant 0. \qquad (1)$$

It is fairly obvious that the addition of equation (1) into Jaffee's basic model[3] will either have no effect on the basic results of that paper or will require that depreciation be taken even later in the time horizon, since that feasibility constraint gives greatest weight to depreciation in the early periods. In this second case, the depreciation path will be even more "decelerated" than Jaffee's Figure 1.[4]

The introduction of both debt and equity financing into the

I would like to thank Wayne Lee for assistance with this comment.

[1] See [2].

[2] See [1].

[3] In [1], p. 339.

[4] In [4], p. 700.

analysis adds realism to the model, but it does exact a price. In addition to the well-known considerations of the optimal debt-equity ratio, there are difficulties in selecting the appropriate discount rate(s). For example, Modigliani and Miller[5] point out that it is generally incorrect to discount cash flow by a single discount rate as is done in Linhart's equation (3).

Finally, the basic results of Linhart's model, which are illustrated in Figure 2, are characterized as representing accelerated depreciation. Whether or not a depreciation schedule can be classified as accelerated requires a comparison of that particular path with straight-line depreciation. We can unambiguously classify a depreciation path as accelerated only if accumulated depreciation at every particular point in time exceeds the sum of past depreciation to that point under a straight line schedule. Thus, at some period betweeen t_0 and T, the optimal depreciation path switches from accelerated to decelerated. The length of the time periods that have accelerated and decelerated depreciation depends upon the values of t_0, A, B, and T.

References

1. JAFFEE, B. L. "Depreciation in a Simple Regulatory Model." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 338–342.
2. LINHART, P. B. "Depreciation in a Simple Regulatory Model: Comment." *The Bell Journal of Economics and Management Science,* Vol. 5, No. 1 (Spring 1974), pp. 229–232.
3. MODIGLIANI, F. AND MILLER, M. "Corporate Income Taxes and the Cost of Capital: A Correction." *The American Economic Review,* Vol. 53, No. 3 (June 1963), pp. 433–443.
4. "Omissio." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 2 (Autumn 1973), p. 700.

---

[5] See [3].

Cable, cities, and copyrights

Bridger M. Mitchell

Senior Staff Economist
The Rand Corporation

in association with

Robert H. Smiley

Assistant Professor
Graduate School of Business
and Public Administration
Cornell University

The Federal Communications Commission has enacted final rules for cable television systems in the 100 largest television markets, restricting cable service to a greater degree than previously proposed. Most systems will be limited to importing only two distant television signals. Furthermore, they are required to "black-out" imported programs that duplicate a local station's showings at any time up to a year before or after they are locally broadcast. Left unresolved is the issue of payment of royalties to copyright owners for the programs carried by cable. This paper examines the profitability of major market cable systems subject to the final FCC rules with particular attention to the effects of several copyright fee proposals, including one incorporated in pending legislation. It uses an earlier model constructed to evaluate the 1970 FCC proposals for cable, with substantial modifications to include both the effects of the new rules and recent research findings on cable demand and cost parameters. The results indicate that, except where atypical conditions prevail, cable systems will be unprofitable in central urban areas having good broadcast reception, but marginally profitable on the edges of the larger cities. In these latter cases rates of return are not far above the estimated cost of capital. However, copyright fees of the magnitude that have been proposed are likely to deter cable construction in many of these areas, since rates of return are sensitive to the additional costs and, because demand is price-elastic, higher prices could not increase revenues.

1. Introduction

■ Final rules of the Federal Communications Commission governing cable television service in the 100 largest television markets went into effect March 31, 1972. The FCC acted after six years of formal proceedings; during that period development of cable

---

We are indebted to Stuart Feldstein, Stanley Besen, and R. E. Park for helpful comments and to Don Andersson for assistance with sources of data. An earlier version of this paper [18] was supported by the National Cable Television Association. The views expressed here are the authors' and do not necessarily reflect those of the National Cable Association, The Rand Corporation, or Cornell University.

service in major cities had been blocked by interim regulations that effectively prohibited the importation of distant television signals. The final rules permit limited importation, varying with the size of the market and the locally receivable signals. At the same time the rules provide broad "exclusivity" protection to local stations for their programs, by requiring cable systems to delete programs under contract to local stations from the imported signals.

The FCC's cable rules have evolved considerably from the 1970 proposals[1] and from Chairman Burch's "letter of intent" to the Chairman of the Senate Subcommittee on Commerce.[2] The final rules are the outcome of a "compromise agreement"[3] among the broadcasting, cable, and copyright interests that was promoted by the President's Office of Telecommunications Policy (OTP). The compromise placed additional restrictions on cable through exclusivity provisions for copyrighted programs shown by local stations, as well as through an agreement by all parties to support legislation that would impose copyright liability for importation of distant signals.

☐ **Signals allowed.** The rules now in effect for the major cities may be summarized as follows.

(1) A cable system must carry all stations licensed within 35 miles of the system and all stations that are "significantly viewed" in the cable system's community.[4]

(2) Cable systems may import distant signals to reach minimum service levels, as follows. In markets of 1 to 50, the minimum service is 3 networks and 3 independents, and programs on imported signals are subject to both one-year preclearance and life-of-contract protection.[5] In markets of 51 to 100, the minimum service is 3 networks and 2 independents, and programs on imported signals are protected for a maximum of two years when under contract to a local station. In markets of 101 or more, the minimum service is 3 networks and 1 independent, and there are no exclusivity limitations.

(3) Cable systems in markets of 1 to 100 may import up to two distant commercial signals after subtracting any signals imported to meet the minimum service standard.

(4) Foreign-language stations may be imported without limit. Noncommercial stations are also unlimited, unless there is local or state objection from educational broadcasters.

---

[1] [32]. An economic analysis of the FCC package appears in [7].

[2] Printed in [5].

[3] In publishing the formal text of the agreement, the FCC preferred the term "consensus," [30], p. 3258.

[4] Network signals are significantly viewed if they have a 3-percent viewing share and 25-percent net weekly circulation; independent signals must have a 2-percent viewing share and 5-percent net weekly circulation. Viewing share indicates roughly the intensity of viewer interest, and net weekly circulation the technical availability of a signal in the community. According to [23], p. 181, 16 of the 37 UHF independents will fail to meet the viewing test in their *own* county.

[5] The exclusivity provisions are discussed in greater detail below.

In the final rules, the FCC has dropped its earlier proposals for a 5-percent tax on cable revenues to support public broadcasting and the scheme to have operators substitute the commercials of local stations for those appearing on imported signals. As previously proposed, the rules contain requirements for 20 channels of capacity in the top 100 markets, a potential for reverse (nonvideo) communication from the home to the cable center, and FCC fees of 30 cents per subscriber. Technical standards are specified, and an annual performance test must be completed.

The new rules significantly increase the requirements for locally programmed channels. For each broadcast channel, the system must have available a nonbroadcast channel. Specifically mandated are three "access" channels for public, educational, and governmental uses. The latter two are to be provided free on a five-year experimental basis; the public access channel must make the first five minutes of time and equipment free to all comers on a nondiscriminatory basis, which ought to increase penetration. The system must have one or more channels available for leasing, and must expand them when demand increases. Finally, the FCC suggests that 3 percent to 5 percent of gross revenues are appropriate levels for local franchise fees, and indicates that higher fees require a showing of cause to the Commission.

Omitted from the rules are any requirements that cable systems pay compensation to copyright owners for the programs carried. The FCC has preferred to leave this issue to the courts and the Congress, and a section of the OTP compromise agreement commits the parties to support legislation setting liability and fees. Thus, apart from copyright fees, the parameters that will affect cable development in the major cities are now determined.

☐ **Simulation model.** In this paper we assess the profitability of cable television in the major markets under the final FCC rules, with particular attention to the effect of several alternative copyright fee schedules that have been proposed. Our research extends the computer simulation model and detailed cost and revenue data developed by Comanor and Mitchell[6] in their study of the impact of the rules proposed in July 1970.

The Comanor-Mitchell model of a cable television system consists of detailed capital and operating cost items specified as functions of system size (the number of subscribers), number and types of signals carried, housing density, type of system construction, and penetration (percentage of potential homes that subscribe). Penetration, or demand, is a function of the price of cable service, family income, and the increase in number and types of television signals provided by the cable system relative to those available off-the-air. Revenue is derived from the system's size, penetration, and price of service, with a small allowance for advertising revenues that may be earned on a local origination channel programmed by the cable system.

The Comanor-Mitchell cost functions were constructed from audited accounting statements, equipment price lists, and *pro forma* estimates supplied by a small sample of system operators. In the original model, the penetration equation was estimated by least squares from a random sample of operating systems. For this paper, however, estimates of penetration are taken from Park's more recent econometric study,[7] as discussed below.

---

[6] [7].

For any given set of market and system characteristics (e.g., number of signals, housing density, system size, etc.), the model simulates the first 15 years of operation of the specified system and calculates the internal rate of return on total capital on the assumption that the system will be rebuilt periodically and generate revenues indefinitely.

In this paper we considerably modify and expand the original model to include the March 1972 FCC rules, more accurate and detailed predictions of penetration in major markets, the effect of the exclusivity provisions on penetration, a comprehensive set of cable system parameters—encompassing market type, available signals, system location, and subscriber and construction characteristics—and four alternative copyright fee schedules (including no fees).

In outline, the analysis of cable profitability focuses on a number of market and system characteristics that can be identified as typical or representative of a cable system if it were to be constructed under current rules. By varying the characteristics (e.g., system size, lineup of local signals, or housing density) over a comprehensive set of possibilities, the outlook for cable in nearly all parts of the major markets can be assessed. In this analysis, costs and prices have been measured in 1970 values; costs, revenues, and rates of return are consequently in "real" terms. Except for rules changes since July 1970, cost figures are based on Comanor and Mitchell's detailed study.

Our analysis includes revenues from subscribers, determined by penetration rates dependent on local and distant signals carried, and a realistic amount from advertising on a local origination channel. No revenues or costs have been attributed to the development of leased channels.[8]

2. The copyright issue

■ Broadcast stations pay copyright royalties to producers of the programs they transmit. Cable systems pay similar fees for movies and other material shown on their local origination channels, and they must also purchase rights to the programming carried on pay or leased channels. But cable systems currently obtain the bulk of their video material—the local and distant broadcast signals—without a payment to either the stations or the copyright holders of their programs.

---

[7] [23].

[8] All systems considered in this study are newly constructed. The effect of potential copyright fees on *existing* systems in comparable market circumstances would be somewhat different only in the short run. For several years, these already built systems would experience reduced profitability and the systems' owners would earn lower returns than they had anticipated. At the same time, revenues would still exceed operating costs, so that the original systems would not actually go out of business. But subsequently, when the systems required rebuilding, the copyright fees could make *reconstruction* unprofitable, since nearly the same investment considerations apply either to rebuilding an existing system or to constructing the same type of system in a similar but unwired community.

The legal status of broadcast signals carried by cable has been based on the *Fortnightly* decisions,[9] in which the Supreme Court held that a cable system acts as a large antenna for the subscriber and does not "perform" the programming material carried over its wires.[10] Consequently, copyright owners have been unable to extract royalties from the cable systems carrying their programs. *Fortnightly* was brought against a system that carried only locally receivable signals and left unresolved the status of signals imported by microwave (or eventually satellite) from distant locations.

In *CBS v. TelePrompTer*,[11] the Supreme Court in March 1974, held that cable systems also incurred no copyright liability by carrying distant signals and said in its opinion, "The reception and rechanneling of signals or simultaneous viewing is essentially a viewer function, irrespective of the distance between the broadcast station and the ultimate viewer."[12] The Court noted that the effects on broadcasters and program suppliers were a matter to be addressed by the Congress.

To assess the economic effects of cable carriage of broadcast signals, we consider first the market for programming in the absence of cable. Broadcast stations earn revenues from selling air time for commercial messages, with advertising rates proportional to the audience for the program on which the commercial appears. The size of the audience a program will receive is determined by its attractiveness relative to that of programs on competing stations and in part by the appeal of programs "adjacent" in time to it on the same station. Networks and independent syndicators compete for the supply of programming created by program producers.[13] If now, a cable system that carries all the local signals is constructed in that community, subscribers have the same menu of signals as before.[14] We may assume that the aggregate audience of the local stations is not diminished by this arrangement, and that advertisers have no reason to pay lower time rates merely because some viewers of their commercials see them over the cable.

When the same system imports independent stations from another community, the audience of the imported stations is enlarged. While total viewing by cable subscribers may well increase because of the greater number and types of program alternatives, viewing of local stations usually declines as well.[15] The first question is to what extent the imported stations' advertising rates will increase to reflect their larger audiences.[16] At present, advertisers on independent stations typically purchase audience "by the market," buying time from one local station in each market they wish to reach. Thus advertisers now ascribe little value to additional viewers outside the station's customary viewing area. But this is not likely to be the long-run pattern. Cable systems will attempt to import the independents with the strongest lineup of programs, subject to the microwave costs of doing so. As cable importation develops, the major independents will enjoy substantial increases in their audiences and will come to be, in effect, regional or national stations.[17] In this form they will compete for regional and national advertising with the existing networks, and rates will reflect their total audience. Advertising revenues of the local network affiliates will decrease while independents will gain revenues. The change will give independents an improved position in bidding for programming.

---

[9] *Fortnightly Corp. v. United Artists Television Inc.*, 392 U.S. 390 (1968).

[10] Liability does attach if the cable system tapes the broadcast program and retransmits it at a later time. See [24], p. 29.

[11] *CBS v. TelePrompTer*, Second Circuit Court of Appeals, Docket No. 72-1800, March 8, 1973, reprinted in [24].

[12] *New York Times*, March 5, 1974.

[13] See, e.g., [3], [10], [20].

[14] For such a system to attract any subscribers, it would either have to improve reception quality or be no more costly than home antennas.

[15] See [12], [20], [22].

Determining the final effects on the distribution and amounts of payments to program suppliers and the supply of programs produced would require a full-scale analysis that we shall not attempt here. Without copyright liability, the long-run development is likely to be a few strong national independent stations whose major audiences are on cable. This configuration has some of the features of a fourth broadcast network, but a different division of revenues. Revenues for such a "network" arise from both advertising sales and cable subscriptions. All advertising income flows to the stations who contract for the programming; all subscriptions are retained by the cable operator.[18] In contrast, regular network affiliated stations carry the network programs and receive a share of the national advertising revenues as well as the revenues of locally sold spot messages they insert into the network programs.

The most likely effect is that there would be some redistribution of compensation from those program owners now supplying the networks to the suppliers of the major independents. New resources might be attracted to the television programming industry. This is particularly likely if cable systems pay copyright fees. But whether the aggregate payments to program owners would increase or decrease without cable copyright liability is an open question. To the extent that networks currently exercise monopsony power in purchasing programming, the rise of a cable network composed of independent stations would reduce the existing networks' power and require them to pay higher rates to suppliers.[19]

☐ **Proposed fees.** In its final cable television rules, the FCC has enacted broad exclusivity protection for copyright holders, yet it also clearly anticipates copyright legislation. In ruling on the cases before them, the courts have indicated that Congressional action is needed. The OTP compromise commits the interested parties to support legislation that establishes, by statute, cable's access to broadcast signals and sets royalty fees for their use. And in response to these concerns, the chairman of the Senate Subcommittee on Patents, Trademarks, and Copyrights, has introduced a bill[20] that would revise the Copyright Law in that manner.

---

[16] Since audience data are routinely collected by county for every station, cable viewing does get attributed to distant stations.

[17] The three largest independents are currently carried on cable systems with total audiences of 1.1 million, 600,000 and 480,000, respectively. *Broadcasting*, July 2, 1973, p. 46. See also [20], p. 170.

[18] The efficiency of this market arrangement in satisfying viewer tastes merits separate analysis.

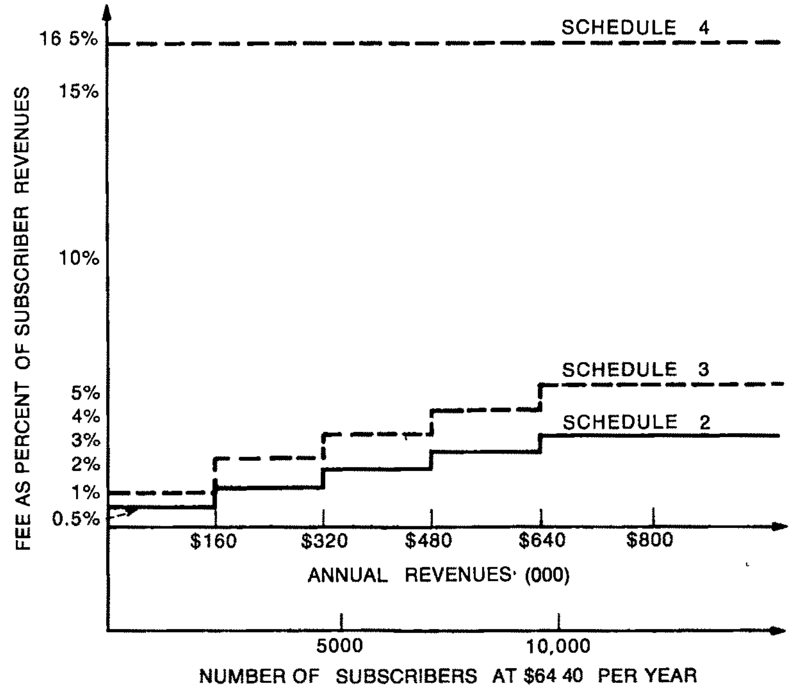
[19] See [10] and [20], Appendix B.

The McClellan bill would affect all cable systems that import distant signals. They would be granted a compulsory license to carry the programming of the imported signals and would make royalty payments on a quarterly basis into a copyright fund. The Register of Copyrights would then disburse the payments to owners of the programming copyrights who may agree among themselves on the division. Disputes are to be adjudicated by a royalty tribunal.

In the following analysis we consider four alternative fee schedules for payment by cable systems to copyright owners. Schedule 1 is the baseline case of zero fees. Schedules 2 and 3 levy successively larger fees as the system's revenue grows. Schedule 3 (incorporated in the McClellan bill) begins at 1 percent of subscriber revenues and rises to 5 percent of revenues exceeding $640,000 annually; Schedule 2 is exactly half of Schedule 3. For Schedule 4, we consider a flat fee of 16.5 percent of subscriber revenues, as proposed by copyright owners during negotiations that followed the OTP compromise agreement.[21] Details of these fees are set forth in Table 1 and in Figure 1.

FIGURE 1

ALTERNATIVE COPYRIGHT FEE SCHEDULES



---

[20] S. 1361, 93rd Congress.

[21] *Broadcasting,* June 25, 1973, p. 39.

TABLE 1

ALTERNATIVE COPYRIGHT FEES

COPYRIGHT FEE SCHEDULE NO. (PERCENT)				ANNUAL SUBSCRIBER REVENUE
1	2	3	4	
0	0.5	1	16.5	OF 1ST $160,000
0	1 0	2	16 5	OF 2ND $160,000
0	1 5	3	16.5	OF 3RD $160,000
0	2 0	4	16.5	OF 4TH $160,000
0	2 5	5	16 5	OF REMAINING REVENUE

☐ **Subscription price.** In comparing systems in different market circumstances and with alternative fee schedules, we keep unchanged the subscriber price as well as the system size and other attributes of cable service. Cable television systems have some of the attributes of a "natural monopoly" because of their high fixed-low variable cost structure. But in practice, the behavior of cable systems is increasingly limited by local and federal regulation and by competition among firms for franchises. Both of these forces sharply restrict the ability of cable firms to adjust price or output at will.

Monthly subscriber rates have remained virtually constant in current prices over several years as a result of these forces plus the threat of more extensive regulatory action if firm behavior is perceived as excessive. In 1970, Seiden[22] found most recently franchised systems charging between $5.00 and $7.00 per month. Comanor and Mitchell reported a mean price of $5.00 per month in their sample of *Television Factbook* systems.[23] Park in 1972 observed an average annual price of $63.[24] The assumption that moderate cost increases, including new copyright fees, cannot be passed on by price increases is consistent with recent market experience.

Assuming no price response by cable firms if a 16.5-percent fee were imposed requires further discussion. Operating firms would doubtless make strong representations to local authorities about the need for higher prices, and bids for new franchises would quote higher rates. But granting for the moment that regulators allowed part or all of the surcharge to be translated into higher subscription rates, how would cable profits be affected? The answer depends primarily on the price elasticity of demand.[25]

Park's study of penetration estimates a nearly unitary elasticity,[26] and independent estimates from a linear probability model applied to individual household data, imply a price elasticity of —0.97 to —1.09.[27] These findings are broadly consistent with the availability of good substitutes for cable television service in urban markets. Households in areas with a diversity of over-the-air signals, with generally clear reception, and with a variety of entertainment alternatives can be expected to decline service rapidly as the monthly price rises.

---

[22] [25].

[23] [7].

[24] [23].

[25] For a discussion of the effect of demand elasticity on maximum rates permitted by a regulatory authority, see [8].

[26] [23], p. 145.

[27] [6].

How, then, would cable systems' profits be affected by a 16.5-percent copyright payment and a concomitant rise in subscriber rates? Revenues would be approximately unchanged, while operating costs would increase sharply by the amount of the copyright payments. Some reductions in other incremental costs would occur by not serving the subscribers who do not purchase service at the higher price. For typical systems, these are the rather small costs of installing additional drop lines, additional maintenance and billing expenses, and higher taxes and dues related to numbers of subscribers. The net effect of allowing higher subscriber rates in conjunction with 16.5-percent copyright fee payments would be to reduce rates of return to nearly the same levels as would be achieved by holding subscriber rates unchanged with the same 16.5-percent copyright fees; and penetration would be lower, providing a narrower base for future leased-channel services.

The discussion in the preceding several paragraphs assumed a degree of upward price adjustment that has not been observed. In the remainder of this study we adhere to a fixed monthly price of \$5.00[28] for maximum cable broadcast service allowed by the FCC rules.[29] An analysis of the profitability of systems under the alternative assumption of higher rates and consequently reduced penetration would yield approximately the same findings.

3. Major parameters

■ **Market characteristics.** In examining the probable effect of various provisions for payment of copyright fees, we consider separately the characteristics of typical cable systems in four types of markets: the top 50 markets, markets ranked 51 to 100, markets above 100, and areas located outside television markets. The FCC rules permit different signal carriage in each situation, and impose differential requirements affecting system costs. In addition, the density of housing, the prevalence of underground utilities, and the level of family income also vary by market size. A summary of these major market characteristics is set forth in Table 2.

As Park's econometric findings[30] demonstrate, the location *within* the market is also of fundamental importance to determining penetration levels. For this study we therefore subdivide each of the markets 1 to 50, 51 to 100, and 100+ into typical "middle market" and "edge market" systems. Middle-market locations are close to off-the-air signals, while edge-market systems are approximately halfway between the transmitter and the B-contour limit

---

[28] Plus \$1.00 for second television sets in 20 percent of the households. No revenue from installation charges is included, as these fees are seldom, if ever, actually collected. See [6] and [7].

[29] Since we are considering all prices and costs in 1970 terms, increases in the monthly subscription rate at about the rate of increase of consumer prices generally will not contradict our observation that *real* subscription rates cannot be adjusted.

[30] [23].

TABLE 2

MAJOR CHARACTERISTICS OF TELEVISION MARKETS

CHARACTERISTICS	LIMITATIONS, BY TYPE OF MARKET			
	1-50	51-100	101+	OUTSIDE
LOCAL SIGNALS TYPICALLY AVAILABLE	3 NETWORKS, OR 3 NETWORKS + 1 INDEPENDENT	3 NETWORKS (SOME UHF)	2-3 NETWORKS (SOME UHF)	1-2 NETWORKS (SOME UHF)
IMPORTED SIGNALS ALLOWED	2 IN ALL CASES, 2-3 INDEPENDENTS TYPICAL	2 IN ALL CASES, 2 INDEPENDENTS TYPICAL	MAXIMUM OF 1 INDEPENDENT PLUS ANY MISSING NETWORKS	NO RESTRICTIONS
EXCLUSIVITY PROTECTION	YES	YES	NO	NO
CAPACITY REQUIREMENTS	20 CHANNELS 2-WAY CAPABILITY	20 CHANNELS 2 -WAY CAPABILITY	12 CHANNELS 2-WAY CAPABILITY	12 CHANNELS 2-WAY CAPABILITY
LOCAL ORIGINATION	STANDARD MINIMUM BELOW 10,000 SUBSCRIBERS	STANDARD MINIMUM BELOW 10,000 SUBSCRIBERS	STANDARD, MINIMUM BELOW 10,000 SUBSCRIBERS	MINIMUM
DENSITY (HOUSES PER MILE)	150-200	125-150	100	80
UNDERGROUND	10-20%	5-10%	5%	0%
FAMILY INCOME (PROJECTED TO 1975)	$11,400-$12,200	$10,000-$11,400	$10,000	$9,000

of the local signals.[31] (The fourth category, an "outside" market system, is necessarily at or beyond the location of a typical edge-market system.) Thus the typical systems analyzed fall into one of seven boxes in the matrix indicated in Table 3.

TABLE 3

CHARACTERISTICS OF TYPICAL SYSTEMS

MARKET TYPE	LOCATION	
	MIDDLE	EDGE
1-50		
51-100		
101+		
OUTSIDE	NOT APPLICABLE	

Within each box, indicating a specific market type/system location, we further consider the two or three most likely lineups of available local signals. While we have not reported every combination that can occur, the cases tabulated represent the major signal patterns to be encountered, and they cover a degree of variation sufficient to include most other possibilities.[32]

☐ **Cable penetration.** Of the many parameters affecting the financial prospects of cable systems, penetration (the percentage of homes passed by the cable that choose to subscribe) is generally the most important. Given that the cable distribution system is already constructed, the marginal cost required to wire an additional home is quite low.

---

[31] Central city areas with reception impeded by tall buildings will have high density but also high penetration rates. *B* contours are calculated from FCC engineering formulas and are intended to enclose areas having satisfactory picture quality 90 percent of the time for 50 percent of the receiving locations.

[32] See Appendix 3, Table A1, for characteristics of signals available in markets 1 to 50 and 51 to 100.

The determinants of cable penetration in fringe reception areas have been investigated statistically by Comanor and Mitchell, by Park, and by Noll, Peck, and McGowan.[33] All three studies use randomly drawn samples of data published by the trade press. While these data introduce some errors of observation, their greatest limitation is that they are drawn almost entirely from systems located in communities with technical reception difficulties, fewer than three network signals, or both.

In contrast, the television environment in the major markets is quite different from this experience. All of the 100 largest markets have three-network service, and 82 of them have all VHF coverage. Thirty-six of the top 50 markets have one or more local independent signals, and three of the markets receive independents from a nearby market that meet the "viewing test." In markets from 51 to 100, seven markets have local independents and six markets receive neighboring independents well enough to meet the viewing test.[34] Equally important, most communities in the top 100 markets have sufficiently high quality signals that they can be received by "rabbit ears" or a nonrotating rooftop antenna. For these reasons, the studies mentioned are of limited value in projecting the cable experience in major cities, as their authors have recognized.[35]

One possible alternative source of projections might be surveys of knowledgeable operators. Unfortunately, this approach has thus far been insufficiently detailed to insure that in their replies the respondents assume the same future state of the world will prevail. In a recent survey[36] of seven large cable firms, which together own 147 separate systems, the questionnaire asked for unconditional forecasts of the number of subscribers, the number of potential subscribers, and capitalization per home at five-year intervals to 1992. No instructions are provided, for example, regarding what regulatory changes should be anticipated, what nonbroadcast services can be assumed, or what will happen to the general price level. Under these conditions it is perhaps not surprising that predictions of cable penetration in 1982 ranged from 31 percent to 70 percent. More disturbing is that the same respondents, when asked what the *actual* industry-wide capitalization per home passed was at the time of the survey, gave answers varying from $60 to $150!

The one possibility for using existing cable experience is to control carefully factors that will differ most in the large urban areas. In a paper published in this *Journal* in 1972, Park carries out the necessary analysis.[37] In doing so, he improves on the earlier studies in three major ways. First, all 63 sample cable systems had at least three good reception quality signals available off-the-air. Systems located in hilly areas or that otherwise had impaired reception were eliminated from the preliminary sample. Second, all data were verified with system operators by telephone interview, insuring greater accuracy than available from only published sources. Third, measures of signal quality were incorporated directly into the analysis. Distance of the cable system from each transmitter was measured, and UHF signals were introduced with a separate parameter to account for more rapid signal attenuation with distance and the absence of UHF tuners in some households.[38]

---

[33] See [7], [22], and [20], Appendix A. Data are taken from [28], [29].

[34] [18], Appendix.

[35] [7], p. 166; [23], p. 131; [20], pp. 290, 292.

[36] Reported in [1].

[37] [23].

The complete penetration equation is estimated by nonlinear least squares and establishes the effects of the following variables:

1. Number of off-the-air VHF signals, by distance from transmitter, with separate categories for networks, duplicate networks, independent, educational, and foreign signals.
2. Number of off-the-air UHF signals, by the same categories, by distance from transmitter, with measurement of UHF set penetration.
3. Number of cable signals, by the same categories.
4. Color set penetration.
5. Annual subscriber price.
6. Annual family income.

Park's study also finds that demand grows to its final penetration level quite rapidly, with subscriptions reaching their equilibrium level about 18 months after a new section of the cable system has first been opened for service. Monthly time series observations of several systems[39] support this finding. In contrast, the growth curves estimated from published data all evidence a slower maturation process.[40] The explanation is that systems with highest penetration were built first, so that in those studies, because quality of reception is not introduced explicitly, the age variable measures the effect of differing reception conditions. The fact that systems are often constructed in several stages has at most a minor effect on penetration estimates. Only when a system is observed less than 18 months after the opening of a new stage of construction will incomplete penetration be observed, and then only for a fraction of the entire system. Therefore, for this study we have adopted Park's estimate curve and assumed that complete penetration is reached in 18 months. Figure 2 compares this growth with that estimated by Comanor-Mitchell.

In the simulations reported below, we project penetration rates using the average figure calculated from Park's equation for each of the major signal characteristics of urban markets. These projected rates can be considered representative of the central experience of systems built in major market communities chosen at random. There is, nevertheless, considerable variation across individual communities in the factors determining penetration. Conse-

---

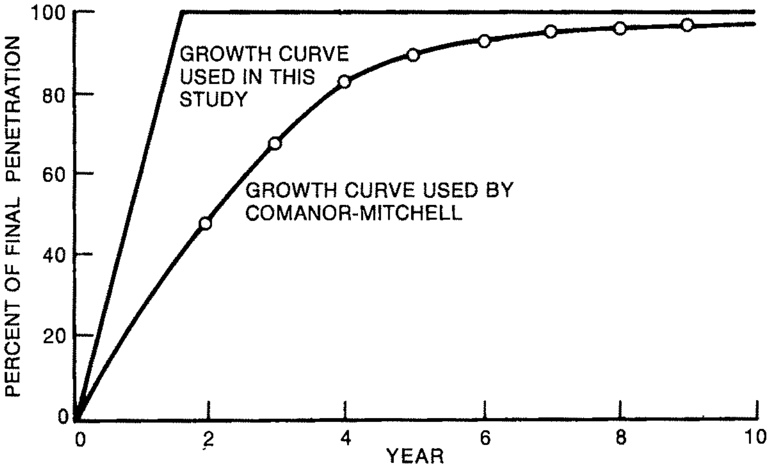
[38] Attempts to test Park's predictions from systems randomly sampled from the *Television Factbook* are invalid, as it is impossible to control for signal quality from published data alone.

[39] Reported in [6].

[40] See [7], [20], [22].

FIGURE 2

RATE OF SUBSCRIBER GROWTH OVER TIME



quently, we also report results for systems using penetration rates 33 percent greater than predicted to prevail under average conditions. Such rates fall in the upper 10-percent confidence interval of Park's equation[41] and are necessarily atypical of the major market environment.

Since investors and operators will seek to build the most profitable systems first, for several years the experience of new systems in the major markets can be expected to resemble that predicted for the most favorable conditions. Even under the quite restrictive 1972 FCC rules and the copyright fee schedule proposed in the McClellan bill, some particularly attractive communities will indeed be wired. But as higher costs or more restrictive signal conditions are imposed, the areas that would otherwise be profitable investments may be rapidly eliminated. Is this the outlook for urban systems as a result of the 1972 Rules, or would it be engendered by significant copyright payments? We shall attempt to answer these questions by reporting the return to investment in systems with average penetration.

☐ **Nonbroadcast services.** Except possibly for a small amount of "local origination" programming produced by individual cable systems, cable television has sold itself by providing more and better-quality broadcast television signals. The technologically more innovative uses of cable's broadband capacity are yet unborn or in extreme infancy. Future video services may eventually include subscription channels or programs for first-run movies, sports that are either not carried on broadcast channels or are locally blacked out, and instructional courses for both the general public and for specialized trade and professional audiences. In addition, the two-way capability of the cable makes possible remote reading of utility meters, detection of fire and burglar alarms and, at a more advanced level, a variety of real-time interactions between home viewers and program originators.

---

[41] [23], pp. 154–156.

Increased gross cable revenues from some of these services are quite likely in the long run. But to date only "pay cable" has been introduced. By the summer of 1973 it was barely under way in 8 systems, with 35,000 homes subscribing.[42] What little experience has been reported indicates that where it has been offered, 10 percent to 30 percent of existing cable homes have signed up for pay channel access to new movies.[43] To date there is almost no evidence as to whether pay channel services increase penetration rates, aside from one report indicating about 1 percentage point increases for two intermediate-sized systems,[44] and pay channels have not yet been introduced in areas where penetration from broadcast signals would be in the range predicted for urban markets.

Prototype business arrangements between system operators, pay channel managers, and program producers take several forms. At one extreme, the three roles may be combined in a single conglomerate firm such as TelePrompTer or Warner Communications. In other instances, the three principal business functions may be performed by independent entities, with the pay cable manager installing and servicing the required terminals and signing up and billing subscribers, but leasing channels from the cable system and purchasing programming from film producers. Various mixed arrangements are also being tried.

Whether new services will make an important contribution to net revenue for cable *operators* is debatable. The FCC requirement that leased channels be available at nondiscriminatory rates provides for entry and considerable potential competition from pay-channel and other nonbroadcast services supplied by other firms.[45] Under these conditions, higher profits for cable operators will accrue only from whatever additional basic service subscriptions may be induced.

In the analysis below we exclude both the additional revenues and costs that would be generated by nonbroadcast services. Only if (1) new services induce important increases in penetration, or (2) cable operators are permitted to earn monopoly returns on such services will rates of return be underestimated.

☐ **Exclusivity restrictions.** The new FCC rules require cable operators to "black-out" numerous classes of programs on imported signals when a local station also shows these programs. These provisions are far broader than the previously announced "simultaneous nonduplication" stricture requiring the cable system to delete from its imported signals any programs that a local station is broadcasting simultaneously. Because the rules also required all local stations to be available on the cable, that provision had no effect on the number of programs available to the viewer.

The new exclusivity rules prohibit cable importation of independent stations' programs to a far greater degree, and because they apply to nonsimultaneous showing, they restrict viewer alternatives. Two sets of exclusivity rules were propounded. In the top 50 markets, cable is automatically prohibited from importing a newly created syndicated program for one year from the date that program is sold anywhere in the country. (The program need *not* be sold to a station in the cable system's market for this "preclearance" provision to be effective.) Thereafter, cable systems are subject to whatever restrictions are included in a local station's contract with the program supplier. Such agreements can prevent importation of a program even though no local station has purchased rights to show it.[46]

---

[42] [17].

[43] *Television Digest*, April 30, 1973

[44] *Broadcasting*, June 25, 1973, p. 32.

[45] [20], p. 194

The exclusivity provisions for markets 51 to 100 are less sweeping. There is no preclearance protection for local stations, and contractual exclusivity arrangements are limited to one to two years, depending on the type of program.

For our purposes, the primary effect of these rules is to reduce the attractiveness of distant signals to subscribers and thus reduce cable penetration. Aside from providing for one channel-switching device for each imported signal, we have not allowed any additional costs of performing the blacking-out function itself, keeping records, etc.

At this writing, evidence on the magnitude of the exclusivity effect is limited to a preliminary study by Park.[47] From detailed program listings for four stations—two networks and two independents—plus partial listings for ten other stations, Park synthesizes the expected proportion of a broadcast week that a distant signal would be blacked out. A portion of his findings is reproduced in Table 4.

TABLE 4

PERCENTAGE OF TIME DISTANT SIGNAL CHANNELS ARE BLACKED OUT

LOCAL SIGNALS	NUMBER OF DISTANT SIGNALS ALLOWED	NUMBER OF DISTANT STATIONS				
		2	3	4	5	6
MARKETS 1-50 3 NETWORK+ 2 INDEPENDENT	2	51	35	26	20	16
3 NETWORK + 1 INDEPENDENT	2	39	24	15	11	8
3 NETWORK	3	52	27	15	9	6
MARKETS 51-100 3 NETWORKS	2	16	6	2	1	0

SOURCE [21], TABLE 2, P. 5.

Park's results indicate, for example, that in those top 50 markets in which local service provides three networks and one independent, the cable system importing two additional independents will be required to black out programs about 39 percent of the time. If it imports a third independent (on a stand-by basis, since the rules allow only two distant signals[48] at any moment on the cable) and "fills in the blanks" where possible, it can reduce the blacked-out time to about 24 percent. Importing a fourth independent further reduces this to 15 percent, etc. The boxed-in figures represent the expected effect when no standby signals are imported.

---

[46] [2].

[47] [21].

The impact of the exclusivity rules on subscriber penetration is likely to be proportionately greater than the reduction in viewing hours. Programs receiving protection will be predominately those with large audiences, many of whom would value an earlier or alternative viewing rate or hour that cable could otherwise provide. Nevertheless, lacking data to make a more refined estimate of this effect, we assume that exclusivity protection is equivalent in its impact on penetration to a proportionate reduction in the number of full-time distant independents carried on the cable, using the appropriate boxed figures from Table 4.

Will it be profitable for a cable system to import standby independent signals? The marginal costs of importing will rise as the cable system must go farther to find each additional independent. Concurrently, the proportion of time that can be filled in with each extra signal is declining. The exclusivity rules thus place the cable firm in a situation of sharply diminishing returns as regards additional penetration from distant signals. Generally, the answer will be "no." Exceptions may occur when the standby independent has particularly attractive programming, or when importation costs depend less on distance, as could occur with satellite transmission.

Regarding importation costs, we have assumed for all systems in this study that distant signals are delivered by cable system-owned microwave links of 50 to 100 miles per channel imported. For several market environments the mean distances to the first and second closest independents (in the top 25 markets) are tabulated in the Appendix. For several types of markets these averages range from 91 to 208 miles to the closest signal, and 125 to 325 miles from the next closest. Thus the microwave cost estimates used here must be considered generally low, although they may be closer approximations for markets with several closely spaced systems that pool their microwave facilities.

☐ **Density and underground construction.** Density, the number of homes per cable mile, can vary considerably from one potential franchise area to another. The Comanor and Mitchell sample reported an average density of 95 per mile inside major markets, and 79 outside.[49] More recently available data for a number of municipalities in the Dayton, Ohio, and Boston, Massachusetts, areas are tabulated in the Appendix. For systems in the study we have assumed somewhat higher densities than Comanor and Mitchell consider, ranging from 80 homes per mile outside of television markets up to 200 homes per mile in the central areas of markets 1 to 50.

---

[48] In a small number of top-50 market communities, there is no local or significantly viewed independent signal. In these areas a third independent may be imported. It must be a UHF station within 200 miles. If none is available, then it must be either a UHF station located anywhere or a VHF station within 200 miles.

[49] [7], p. 177.

In practice, of course, both higher and lower densities will be encountered. But the tendency to a substantially higher figure for any important number of similar systems is unlikely in view of the FCC's emphasis that it will not authorize carriage of broadcast signals by systems that do not serve all parts of the community.[50]

When part of the cable system must be placed underground, rather than strung from existing utility poles, construction costs increase substantially. Furthermore, the cost per mile varies greatly according to the material that must be trenched and the protection the cable must receive.[51] In some instances existing utility ducts have excess capacity and most digging can be avoided by paying utility companies to install the cable and for space rented.

We have allowed for 5 to 20 percent of the cable system to be placed underground in the major markets. In doing so, we have assumed that municipal authorities will require the same standards of construction they now apply to their telephone and electric utilities. The range adopted here is consistent with the data reported in the Appendix.

4. Results

■ **Measuring cable profitability.** To summarize the profitability of the typical cable systems of this study, we shall calculate the (pretax) internal rate of return on total capital invested in each system. The rate of return required to induce investment in a system will depend on the proportion of total capital that can be obtained through debt instruments, the borrowing rate, and the minimum return demanded by equity investors. One method of measuring the required return would be to compute the cost of equity capital to the cable television industry according to the capital asset pricing approach.[52] To be correct, such a calculation must construct a portfolio of cable firms that are members of the same equivalent-risk class, take into account differences in firm leverage,[53] and be adjusted for regression bias[54] and expected inflation rates. When combined with a long-run estimate of the debt-equity ratio for such firms, a cost of total capital may be developed.

Another approach, and the one adopted here, is to rely on the recent experience of cable firms in obtaining new financing. A detailed 1971 investment survey[55] of the cable industry reports that mature companies with demonstrated earnings have obtained bank credit lines and revolving credit arrangements at 0.5 percent to 1 percent over the prime rate, with long-term debt at 3 percent to 5 percent above prime. Sources of long-term financing are reported looking for a 15-percent return on combined debt-equity instruments. At a more recent industry discussion of financing for cable systems,[56] a major intermediary reported arranging equity financing at around 14 percent.

---

[50] *Federal Register*, p. 3276, §180.

[51] See [7], Appendix 2.

[52] [26].

[53] [14].

[54] [4]

[55] [11].

Consumer prices increased at an annual rate of nearly 4 percent during 1970–1972, accelerating to an 8-percent rate in the first half of 1973. Presuming that capital markets were reflecting an expected 4- to 5-percent rate of inflation, a 15-percent rate of return on total capital corresponds to a 10- to 11-percent cost of capital for cable systems in real terms.

For this study, we have held both revenues and costs at 1970 price levels over the full life of the cable system.[57] Financial measures are therefore in real terms, and are appropriately compared to the cost of capital net of changes in the general price level. We assume that the full cable system is constructed in the initial year of operation. Alternatively, one could allow for a series of construction stages with correspondingly staggered demand curves for each stage.[58] Because only a small portion of the cable system capital is completely "fixed,"[59] the calculated rate of return would be nearly the same.

☐ **Example.** We are now prepared to analyze the financial results for typical systems in the several market situations discussed earlier. The detailed cost and revenue schedules built into the Comanor-Mitchell computer program have been modified to include the changes in FCC rules, penetration, and costs discussed above and in the Appendix. Using these data and parameters for the specific system, the computer simulates the complete revenue and cost experience to be expected.

As an example, consider a representative system of 25,000 subscribers located near the middle of one of the 50 largest markets, as reported in the first line of Table 5. Density is assumed to be 200 homes per mile, and family income $12,200.[60] Since this is a central urban location, 20 percent of the cable miles are underground. Annual subscriber rates are $62.40, corresponding to $5.00 per month plus a small additional amount for second sets. We assume that systems of this size will operate with "standard" local origination equipment and will derive modest revenues from selling advertising on the cablecasting channel, estimated at $2.20 per subscriber annually.

The table shows that three VHF networks plus one viewing-test network are available off-the-air. Also, there is one UHF independent and one VHF educational station. In addition to these

---

[56] *Broadcasting*, June 25, 1973, p. 34.

[57] The Comanor-Mitchell specification provides for some wage increases during the initial years of employment to reflect low entering wages and increasing job productivity [7], pp. 202–203. Weinberg's very detailed cost study of cable system components is in close agreement with the system costs used here [33].

[58] See, e.g. [15].

[59] Largely the tower and head-end equipment.

[60] The effect of rising real incomes has been introduced by projecting final penetration based on mean real family income in 1975 for each type of market. Income figures were derived by assuming a 2-percent real rate of increase from 1970 data.

TABLE 5

CABLE PROSPECTS IN MIDDLE MARKETS

SIGNALS[a]					LARGE SYSTEMS INTERNAL RATE OF RETURN (15 YEAR LIFE)				INTERMEDIATE SYSTEMS INTERNAL RATE OF RETURN ,(15 YEAR LIFE)			
LOCAL	VIEWING TEST	IMPORTED BY CABLE	FAMILY INCOME	PENETRATION	COPYRIGHT FEE SCHEDULE NO 1	2	3	4	COPY FEE SCHEDULE NO 1	2	3	4
MARKETS 1-50, DENSITY =200, UNDERGROUND=20%					SIZE = 25,000				SIZE = 10,000			
1 3NV,1IU,1EV	1NV	2I,1E	$12,200	27 2%	10 3%	9 6%	8 8%	4 1%	5 6%	5 1%	4 7%	*
2 3NV,1EU	–	3I,1E	11,400	24 7	8 8	8 1	7 4	2 6	4 3	3 8	3 4	*
3 3NV,1IV,1IU, 1EV	–	1I,1E	12,200	21 9	6 7	6 0	5 3	0 5	2 6	2 2	1 7	*
MARKETS 51-100, DENSITY = 150, UNDERGROUND = 10%					SIZE = 25,000				SIZE = 10,000			
4 3NV, 1EU	–	2I,1E	$11,400	24 0%	4 9%	4 2%	3 5%	*	0 2%	0 1%	*	*
5 2NV, 1NU	–	2I,1E	10,200	30 7	8 0	7 3	6 6	1 8	3 6	3 1	2 6	*
6 1NV, 2NU	–	2I,1E	10,000	35 3	10 4	9 7	9 0	.4 0	5 9	5 5	5 0	*
MARKETS 101+, DENSITY=125, UNDERGROUND=5%					SIZE = 15,000				SIZE =7,500[b]			
7 1NU, 2NU	–	1I,1E	$10,000	29 6%	9 7%	9 0%	8 4%	2 0%	6 5%	6 1%	5 7%	*
8 2NU	–	1N,1I,1E	10,000	55 4	24 0	23 1	22 3	15 0	19 9	19 4	18 9	11 1

[a] IN THESE TABLES, N MEANS NETWORK, I MEANS INDEPENDENT, E MEANS EDUCATIONAL, V MEANS VHF, U MEANS UHF

[b] MINIMAL ORGANIZATION FACILITIES AND NO ADVERTISING REVENUES

*NEGATIVE RATE OF RETURN

broadcast signals, the cable system imports two independents and one educational station. We assume these signals are imported by microwave, averaging three "hops" of 35 miles each per channel.

Within two years the system is assumed to reach maturity, apart from further growth due to enlargement of its franchise area. Penetration is predicted to be 28.1 percent if the distant signals are fully available, but 27.2 percent as a result of exclusivity protection on the independent channels.

In the remainder of the table we assess the effect on profitability of each of the four copyright fee schedules. If fixed capital equipment is replaced about every 15 years, this system will earn a 10.3-percent real rate of return on total invested capital, absent any copyright fees. Alternatively, the proposed statutory schedule (number 3) reduces the rate of return to 8.8 percent, and the flat 16.5-percent fee lowers returns sharply to 4.1 percent. Assuming a 10-year lifetime for equipment would reduce these returns by about 3 percentage points.

In the analysis below we report rates of return based only on 15-year lifetimes. Fifteen years represent a compromise between somewhat longer physical lifetimes for some parts of the cable plant and rather shorter economic lifetimes of currently operating systems experiencing technological obsolescence. The rate of recent innovation in electronics suggests that cable systems will continue to undergo major changes.

☐ **Detailed results.** The financial prospects for cable under the final FCC rules and the effect of alternative copyright fee schedules are contained in Tables 5, 6, and 7. While we shall briefly review the major findings, the reader should consult the tabulations for particulars.

Table 5 reports the expected experience in *middle markets* of large- and intermediate-sized systems. Line 1 of the table has been discussed in detail in the example above. Lines 2 and 3 are for similarly situated communities with somewhat different sets of local signals. Penetration ranges from 22 to 27 percent and rates of return from 6.7 to 10.3 percent for large systems when there are no copyright fees. Despite somewhat higher penetration rates, systems in the second 50 middle markets earn generally lower returns, principally because of reduced density, while in the lowest ranked markets there is great variation, with profitable 55-percent penetration systems when one network is missing from the local signals.

Intermediate-sized systems in middle markets are decidedly below the 10-percent rate of return needed to attract investment funds. Except where quite large systems of 25,000 or more subscribers can be built, central city areas of the major markets are not bright prospects for cable under present rules, even without copyright payments.

The prospects for large systems at the *edge* of major markets[61] in Table 4 are brighter. In the top 50 markets, penetration is in the 34- to 38-percent range, with rates of return between 11.4 and 13.3 percent. In the second 50 markets, penetration ranges up to 45 percent, with rates of return from 10.0 to 14.6 percent. In the smallest markets and also the fringe (outside) areas, we find more heterogeneous results, with quite profitable cable possibilities where fewer than three networks are available.

The corresponding intermediate-sized edge systems are again unprofitable in all three network cases. While large systems should be feasible in the major metropolitan areas, as of February 1973 only 25 systems had more than 20,000 subscribers, and the largest had less than 60,000.[62] Some fraction of these economies of scale can be achieved when a series of smaller systems is under common ownership, thereby realizing savings from efficient use of management and technical personnel and from sharing local programming and signal importation expenses.

The results presented in Tables 5 and 6 are based on market, economic, and construction factors typifying the most common situations that will be encountered in middle and edge locations of each of the four types of markets. Of course, within each category there will be a degree of variation, clustered around the typical situations we report. Some communities will have higher incomes, others will require extensive undergrounding, still others will require elaborate local origination facilities, etc.

To measure the sensitivity of our findings for typical systems to such variations, we have rerun all of the intermediate-sized systems assuming that penetration is *one-third greater* than would be expected, on average, for each set of market characteristics. A variety of unmeasured factors can cause actual penetration to vary above or below the average value predicted by the penetration equation. In increasing the average value by one-third, we have, in effect, selected only the *10 percent* of the cases in which penetration is *most favorable*; nine out of ten communities having the same signal lineups, income, etc., will have lower penetration.

---

[61] Roughly 35 miles or more from the city center.

[62] *Broadcasting*, May 28, 1973, p. 53.

TABLE 6

CABLE PROSPECTS IN EDGE MARKETS

SIGNALS [a]					LARGE SYSTEMS INTERNAL RATE OF RETURN (15 YEAR LIFE)				INTERMEDIATE SYSTEMS INTERNAL RATE OF RETURN (15 YEAR LIFE)			
					COPYRIGHT FEE SCHEDULE NO				COPYRIGHT FEE SCHEDULE NO			
LOCAL	VIEWING TEST	IMPORTED BY CABLE	FAMILY INCOME	PENETRATION	1	2	3	4	1	2	3	4
MARKETS 1-50, DENSITY = 150, UNDERGROUND = 10%					SIZE = 25,000				SIZE = 10,000			
1 3NV,1IU,1EV	1NV	2I,1E	$12,200	37 9%	13 3%	12 6%	11 8%	6 5%	7 8%	7 3%	6 8%	*
2 3NV,1EU	—	3I,1E	11,400	37 1	13 0	12 2	11 5	6 2	7 5	7 0	6 5	*
3 3NV,1IV,1IU, 1EV	—	1I,1E	12,200	34 4	11 4	10 6	9 9	4 7	6 4	5 9	5 4	*
MARKETS 51-100, DENSITY = 125, UNDERGROUND = 5%					SIZE = 25,000				SIZE = 10,000			
4 3NV,1EU	—	2I,1E	$11,400	36 3%	10 0%	9 3%	8 5%	3 2%	5 4%	4 9%	4 4%	*
5 2NV,1NU	—	2I,1E	10,200	38 4	11 5	10 7	10 0	4 6	6 7	6 2	5 7	*
6 1NV,2NU	1NV	2I,1E	10,000	44 9	14 6	13 8	13 0	7 6	9 1	8 6	8 1	1 1
MARKETS 101+ DENSITY = 100, UNDERGROUND = 5%					SIZE = 15,000				SIZE = 7500 [b]			
7 1NV,2NU	—	1I,1E	$10,000	41 5%	12 2%	11 5%	10 8%	4 2%	8 7%	8 3%	7 9%	0 3%
8 2NU	—	1N,1I,1E	10,000	65 5	23 3	22 5	21 7	14 4	19 3	18 9	18 4	10 5
OUTSIDE MARKETS DENSITY = 80, UNDERGROUND = 0%					SIZE = 15,000				SIZE = 7500 [b]			
9 2NV	—	1N,3I,1E	$9,000	50 1%	13 4%	12 7%	11 9%	5 4%	8 6%	8 1%	7 7%	*
10 2NU	—	1N,3I,1E	9,000	65 5	20 2	19 4	18 6	11 7	15 5	15 1	14 6	7 0

[a] IN THESE TABLES, N MEANS NETWORK, I MEANS INDEPENDENT, E MEANS EDUCATIONAL, V MEANS VHF, U MEANS UHF

[b] MINIMAL ORIGINATION FACILITIES AND NO ADVERTISING REVENUES.

* NEGATIVE RATE OF RETURN.

Turning to the results in Table 7, we find that such unusually high penetration is sufficient to produce at least one profitable system in each type of market, at least if copyright fees are absent. We conclude that 7,500 to 10,000 subscriber systems can expect to earn a going rate of return in the top 100 markets only when local circumstances are unusually favorable.

Finally, we may assess the financial prospects for cable when copyright fees are required. The predominant effect of Schedule 3, the statutory fees proposed in the McClellan bill, is to reduce the rate of return on total capital 1.5 percentage points for profitable and near profitable systems, and by somewhat less for systems well below the 10-percent return level. Thus, in the example system (the first line of Table 5), the rate of return falls from 10.3 to 8.8 percent.

A one-point change in the real rate of return on total capital has a considerably larger effect on equity holders. If one-half to two-thirds of the cable system is financed by 6-percent[63] debt instruments, this leverage converts a 10-percent return on total capital to a return on equity of up to 14 or 18 percent. In consequence, a decline to a 9-percent return on total capital can reduce the return on equity by 2 to 3 percentage points, depending on the capital structure of the system. Changes of this magnitude are more than sufficient to postpone or eliminate construction of cable systems that otherwise appear marginally profitable.

---

[63] In an inflationary period, borrowing costs would be higher by approximately the expected rate of inflation.

TABLE 7

CABLE PROSPECTS FOR INTERMEDIATE SYSTEMS UNDER 10% MOST FAVORABLE PENETRATION CONDITIONS

SIGNALS[a]				IN MIDDLE MARKETS					IN EDGE MARKETS				
LOCAL	VIEWING TEST	IMPORTED BY CABLE	FAMILY INCOME	PENETRATION	INTERNAL RATE OF RETURN (15 YEAR LIFE) COPYRIGHT FEE SCHEDULE NO				PENETRATION	INTERNAL RATE OF RETURN (15 YEAR LIFE) COPYRIGHT FEE SCHEDULE NO			
					1	2	3	4		1	2	3	4
MARKETS 1-50 , SIZE = 10,000				DENSITY = 200, UNDERGROUND = 20%					DENSITY = 150, UNDERGROUND = 10%				
1. 3NV,1IU,1EV	1NV	2I,1E	$12,200	36 3%	10 8%	10 3%	9 9%	3 7%	50 5%	13 3%	12 8%	12 3%	5 7%
2. 3NV,1EU	—	3I,1E	11,400	32 9	8 1	7 6	7 2	0 7	49 5	13 1	12 6	12 1	5 5
3 3NV,1IV,1IU 1EV	—	1I,1E	12,200	29 2	6 5	6 1	5 6	*	45 9	11 9	11 4	10 9	4 4
MARKETS 51-100, SIZE = 10,000				DENSITY = 150, UNDERGROUND = 10%					DENSITY = 125, UNDERGROUND = 5%				
4 3NV,1EU	—	2I,1E	$11,400	32 0%	5 3%	4.8%	4 4%	*	48 4%	9 7%	9 2%	8 6%	1 6%
5 2NV,1NU	—	2I,1E	10,200	40 9	7 6	7 1	6 6	*	51 2	11 1	10 5	10 0	2 9
6. 1NV,2NU	—	2I,1E	10,000	47 1	11 4	10.9	10 5	4 2	59 9	15 0	14 4	13 9	7 0
MARKETS 101+, SIZE = 7500[b]				DENSITY = 125, UNDERGROUND = 5%					DENSITY = 100, UNDERGROUND = 5%				
7. 1NV,2NU	—	1I,1E	$10,000	49 5%	17 7%	17 2%	16 8%	9 2%	55 3%	16 1%	15 6%	15 2%	7 8%
8 2NU	—	1N,1I,1E	10,000	73 9	25 5	25 0	24 4	15 7	87 4	25 0	24 4	23 9	15 2
OUTSIDE MARKETS SIZE = 7500[b]				NOT APPLICABLE					DENSITY = 80, UNDERGROUND = 0%				
9. 2NV	—	1N,3I,1E	$9,000						66 9%	15 8%	15 4%	14 9%	7 3%
10 2NU	—	1N,3I,1E	9,000						87 3	20 7	20 2	19 7	11 5

[a] IN THESE TABLES, N MEANS NETWORK, I MEANS INDEPENDENT, E MEANS EDUCATIONAL, V MEANS VHF, U MEANS UHF.

[b] MINIMAL ORIGINATION FACILITIES AND NO ADVERTISING REVENUES.

The preponderance of evidence in Tables 5, 6, and 7 is that large systems at the edges of the top 100 markets will earn a 10- to 14-percent rate of return before copyright payments, large systems in middle markets are not likely to exceed 10 percent, and intermediate and smaller-sized systems will be marginally profitable only where special factors operate. Copyright fees, at the level of Schedule 3, would significantly slow the growth rate of cable in the major markets, particularly in middle areas with good quality signals and in edge market communities of intermediate size.

Copyright fee Schedule 2 is exactly one-half the rate of Schedule 3. As expected, it has approximately half the effect of Schedule 3 in reducing the rate of return for all systems.

Schedule 4 is the flat 16.5-percent copyright fee. Its effect on rates of return is devastating. Of all variations studied in the top 100 markets, only systems extrapolated to a 50,000 subscriber level at the edge of markets 51 to 100 reach a 10-percent rate of return. Fee payments of this magnitude would effectively halt cable growth in the large cities.

5. Conclusion

■ The outlook for early development of cable television service in the major cities is at best mixed. As compared with the rules discussed two years ago,[64] the final FCC rules more tightly restrict the number and choice of broadcast signals a system can provide

---

[64] See [7].

to its subscribers and include exclusivity provisions that require deleting some portion of those programs.

Our analysis of the important variations in potential market and cable system characteristics in these urban areas demonstrates that only the largest systems, or multiply-owned systems of slightly smaller scale, will be viable in the central city areas where off-the-air reception quality is high, and then only under favorable construction and penetration conditions. At the edges of these markets, returns will be sufficient to attract investment in the largest (25,000 subscriber) systems, but systems of 10,000 to 15,000 subscribers will be profitable only under especially favorable circumstances.

In an investment environment that allows most urban households to be profitably wired for cable television service only when atypically propitious cost and demand factors occur, requiring more than quite limited copyright payments will significantly retard or halt cable expansion in the urban markets. The proposed statutory fee schedule in the McClellan bill (up to 5 percent of subscriber revenue) would generally lower rates of return on total capital 1.5 percentage points for systems in the profitable range; and in an important proportion of cases, its leveraged effect on equity investors would be sufficient to create unprofitable systems.

As expected, a fee schedule of one-half that proposed reduces rates of return on total capital somewhat more than one-half a percentage point. Fees of this magnitude would restrict cable construction primarily in market circumstances in which returns are already limited for other reasons. In contrast, a flat 16.5-percent copyright payment would create a decidedly unprofitable investment climate for cable television throughout the top 100 markets, far outweighing the limited prospects opened up by the 1972 FCC rules.

Appendix 1

Modified costs and revenues

■ Several cost items in the Comanor-Mitchell study have been modified for this study, either to take account of the FCC rules as finally adopted or as a result of the availability of more recent information. A brief summary of those costs which were modified for *all* systems investigated in this report is presented below.

(1) *Local franchise tax.* 5 percent of gross revenues annually.
(2) *FCC fee.* $35 initial fee plus $0.30 per subscriber annually.
(3) *Channel switchers.* One switcher included in capital equipment costs for each imported signal.
(4) *Pole rent.* All results reported here include pole rent of $250 per aerial mile in top 100 markets, $175 in other markets.
(5) *Local origination.* We assume the Comanor-Mitchell *standard systems*, with capital costs of $38,000 and annual operating expenses of $4,300, and for smaller systems a *minimum system*, with capital costs of $11,000 and operating expenses of $2,500 per year for live origination. All

systems are assumed to provide a time-and-weather channel.

(6) *Public service channels.* The final FCC rules require CATV systems to provide 3 nonbroadcast channels for noncommercial public, educational, and government access, respectively. The public access channel is to be provided without charge, while the other two channels will be free for five years. The costs of meeting these provisions are taken to be an additional 75 percent of the capital costs assumed for local origination, plus $4,875 per year for part-time technicians' salaries.

(7) The previously proposed 5-percent "public dividend" tax for support of noncommercial broadcasting has been eliminated.

(8) *Rate of subscribers' growth over time.* Two recent studies using unpublished data[65] indicate a more rapid rate of growth than was assumed in an earlier version of this paper.[66] We have revised those calculations and based the figures reported here on the assumption that the cable system reaches final penetration after 18 months. Revenues and operating costs are discounted at the end of each 12-month period following system construction, so that net revenues are discounted from the midpoint of the first "year" in which penetration reaches its final level, and similarly thereafter.

As in the Comanor-Mitchell paper, internal rates of return are calculated for a firm of infinite life. We assume that the firm reaches an equilibrium of revenues and costs after 18 months and that the lifetime of the system's physical assets is 15 years. Thereafter, the plant is rebuilt periodically, while subscriber penetration is held constant at the mature level. The rate of return is generally robust with respect to exact assumptions about conditions in later generations. Another solution to this terminal value problem is to assign the firm a value at the end of its first generation, based on operating characteristics such as revenues, subscribers, etc.[67]

Appendix 2

The penetration equation

■ Technical details of the penetration equation are summarized below. For further discussion see Park's earlier study.[68]

$$\log\left(\frac{Pen}{1-Pen}\right) = -8.159 + 3.098 \log X_N + 0.290 \log X_D + 0.212 \log X_I + 0.298 \log X_E - 0.540 \log X_F - 1.473 \log P + 1.398 \log Y + 0.523 \log C,$$

where

---

[65] [23], [6].

[66] [18].

[67] For an example of this method, see [15].

[68] In [23].

$$X_i = \frac{1 + W_i}{1 + 0.731 U \sum_{U_i} (1 - d^{1.6})^{1/1.6} + \sum_{V_i} (1 - d^{1.6})^{1/1.6}}$$

$i = N =$ network
$D =$ duplicating network
$I =$ independent
$E =$ educational
$F =$ foreign
$W_i =$ number of cable signals of type i
$U_i =$ number of B-contour off-air UHF signals of type i
$V_i =$ number of B-contour off-air VHF signals of type i
$Pen =$ penetration $=$ subscribers/households passed by cable
$P =$ annual price
$Y =$ median family income
$C =$ color set penetration
$U =$ UHF set penetration.

In order to use Park's estimated equation to predict penetration for the typical systems investigated here, in all of the simulations reported we assign representative values to the variables as follows:

$P =$ \$62.40, corresponding to the \$5 per month plus \$1 per month for 20 percent of subscribers as a charge for second set.
$C =$ 50 percent. The effect of varying color set penetration is not estimated with sufficient precision to incorporate variations in color set ownership across different types of markets.[69]
$U =$ 80 percent if 0 local network UHF signals
90 percent if 1 local network UHF signals
95 percent if 2 local network UHF signals
99 percent if 3 local network UHF signals.
$F = 0$. Foreign stations are not included among the signals carried by study systems.

In simulating cable systems for this study, we consider systems located in the central area of a television market, where off-the-air signal quality is generally high, and outlying areas of the same market, where quality is diminished. In the penetration equation the distance variable d is a measure of the reduction in quality. A d value of 0 corresponds to a viewer in the center of the market, while a value of 1 represents a viewer at the B-contour of the off-the-air signal.

For the systems in this study we have used the following values for VHF stations:

In *middle markets*:
$d = 0$ for local stations
$d = 1$ for viewing-test stations.
In *edge markets*:
$d = 0.5$ for local stations
$d = 0.75$ for viewing-test stations.

---

[69] [23], p. 145.

The prospects for cable under the 10-percent most favorable penetration conditions are calculated using 133 percent of the penetration implied by Park's equation above. This corresponds approximately to the penetration value at the upper 10-percent confidence limit.[70]

Appendix 3

■ The following tables indicate the characteristics of signals available in the top 100 markets, the households, population, and density in Dayton, Ohio, and the housing density and underground conditions for electric utility service in the Boston, Massachusetts, area.

TABLE A1

AVERAGE CHARACTERISTICS OF SIGNALS AVAILABLE IN TOP 100 MARKETS BY TYPE OF LOCAL NETWORK SIGNALS

LINE	NUMBER OF MARKETS	LOCAL NETWORK SIGNALS	INDEPENDENT		EDUCATIONAL		VIEWING TEST		DISTANCE TO CLOSEST TOP 100 INDEPENDENTS		FAMILY INCOME (1975)
			V	U	V	U	NET	IND	1st	2nd·	
MARKETS 1-50											
(1)	34	3 NV	0 6	1 4	0 6	0 6	0 1	0	208	325	$12,229
(2)	11	3 NV	0	0	0 4	0 5	0 3	0 2	184	252	11,416
MARKETS 51-100											
(3)	21	3 NV	0	0	0 4	0 4	0	0	198	325	$11,371
(4)	7	2 NV, 1NU	0	0	0 1	0 4	1 1	0 1	163	252	10,217
(5)	5	1 NV, 2NU	0	0	0 2	0 2	0	0	91	125	11,713
(6)	6	3 NU	0	0 2	0	0 3	0	0 2	114	174	11,611

SOURCE · [19], APPENDIX, COMPUTED FROM [28], PP 51A-56A

TABLE A2

HOUSEHOLDS, POPULATION AND DENSITY IN DAYTON, OHIO (MARKET NO. 41)

A DAYTON URBAN AREA			
CITY	POPULATION (1970)	DWELLING UNITS (1970)	DENSITY (DWELLING UNITS PER STREET MILE)
DAYTON	243,601	85,401	149
KETTERING	69,599	22,809 .	110
FAIRBORN	32,267	10,156	107
MIAMISBURG	14,797	4,839	71
VANDALIA	10,796	3,335	96
WEST CARROLLTON	10,748	3,476	87
CENTERVILLE	10,333	2,984	66
OAKLAND	10,095	3,795	90
FOUR COMMUNITIES LESS THAN 10,000 PERSONS	23,364	7,683	—
TOTAL	425,620	144,487	119
B DAYTON URBANIZED AREA (EXCLUDING URBAN AREA ABOVE)			
	583,000	192,000	110

SOURCE [15]

---

[70] *Ibid.*, pp. 145–146.

TABLE A3

HOUSING DENSITY AND UNDERGROUND CONDITIONS FOR ELECTRIC UTILITY SERVICE IN THE BOSTON, MASSACHUSETTS AREA

COMMUNITIES SERVED BY BOSTON EDISON COMPANY				
CITY OR TOWN	POPULATION	HOUSING UNITS	DENSITY	UNDER-GROUND
BOSTON	641,071	232,448	324	63 2%
ACTON	14,770	4,195	49	0 4
ARLINGTON	53,524	17,921	143	5 8
BEDFORD	13,513	3,327	55	6 1
BROOKLINE	58,886	23,323	250	47 7
BURLINGTON	21,980	5,423	70	1 3
CANTON	17,100	4,673	55	7 7
CHELSEA	30,625	10,494	238	18 7
DEDHAM	26,938	7,776	78	16 2
FRAMINGHAM	64,048	19,451	95	3 3
HOLLISTON	12,069	3,248	43	0 9
LEXINGTON	31,886	8,859	58	8 5
MILTON	27,190	8,058	77	7 1
NATICK	31,057	8,984	65	12 8
NEEDHAM	29,748	8,674	65	5 8
NEWTON	91,066	27,425	92	11 0
SHARON	12,367	3,470	36	3 7
SOMERVILLE	88,779	29,757	287	23 5
STONEHAM	20,725	6,514	90	8 2
SUDBURY	13,506	3,394	31	12 3
WALPOLE	18,149	4,838	57	12 4
WALTHAM	61,582	18,560	119	8 9
WATERTOWN	39,307	12,842	169	13 7
WAYLAND	13,461	3,669	40	2 7
WESTON	10,870	2,785	34	1.9
WESTWOOD	12,750	3,570	51	5.8
WINCHESTER	22,269	6,543	89	11 1
WOBURN	37,406	10,794	95	7 9
ELEVEN TOWNS OF LESS THAN 10,000 POPULATION	70,950	18,477	39	1 9
TOTAL.	1,587,592	519,492	141	19 0

SOURCE
1) POPULATION AND HOUSING UNITS FROM THE 1970 CENSUS
2) DENSITY AND UNDERGROUND COMPUTED FROM MILES OF STREETS WITH EDISON LINES IN "STATISTICS ON CITIES AND TOWNS SERVED BY BOSTON EDISON COMPANY ENGINEERING AND CONSTRUCTION DEPARTMENT - RECORD DIVISION," DECEMBER 31,1972

References

1. Andersson, D. *The CATV Industry: Current Subscriber Penetration, Projected Subscriber Growth, Capitalization, by 5-Year Periods, 1972–1992.* National Cable Television Association, September 1972.
2. Besen, S. M. "The Economics of the Cable Television 'Consensus'." *Journal of Law and Economics,* forthcoming.
3. ——— and Soligo, R. "The Economics of the Network-Affiliate Relationship in the Television Broadcasting Industry." *The American Economic Review,* Vol. 63, No. 3 (June 1973), pp. 259–268.
4. Blume, M. E., II, "On the Assessment of Risk." *Journal of Finance,* Vol. 26, No. 1 (March 1971), pp. 1–10.
5. Burch, D. "FCC Plans for CATV Regulation," in Television Digest, Inc., *CATV and Station Coverage Atlas and 35-Mile Zone Maps,* 1971–1972, Washington, D. C., 1971.
6. Charles River Associates, Inc. *Analysis of the Demand for Cable Television.* Report No. 78-2, prepared for Office of Telecommunications, U. S. Department of Commerce, April 1973.
7. Comanor, W. S. and Mitchell, B. M. "Cable Television and the Impact of Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 154–212.

8. ———. "The Costs of Planning: The FCC and Cable Television." *Journal of Law and Economics*, Vol. 15, No. 1 (April 1972), pp. 177–206.
9. ———. "The Lost Generation: A Correction." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 704–705.
10. CRANDALL, R. W. "FCC Regulation, Monopsony, and Network Television Program Costs." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 483–508.
11. EQUITY RESEARCH ASSOCIATES. *The Cable Television Industry*. New York: Halle & Stieglitz, Inc., October 19, 1971.
12. FISHER, F. M., BELSLEY, D., FERRALL, V. E., AND MITCHELL, B. M. "Community Antenna Television Systems and Local Television Station Audience." *Quarterly Journal of Economics*, Vol. 80 (May 1966), pp. 227–251.
13. FOUNDATION 70. *Cable in Embryo: Economic Considerations for Urban Franchising*. Wellesley, Mass.: September 1971.
14. HAMADA, R. S. "The Effect of the Firm's Capital Structure on the Systematic Risk of Common Stocks." *Journal of Finance*, Vol. 27, No 2 (May 1972), pp. 435–452.
15. JOHNSON, L. L., BAER, W. S., BRETZ, R., CAMPH, D., FELDMAN, N. E., PARK, R. E., AND YIN, K. K. *Dayton Miami Valley: Basic Report*. The Rand Corporation, R-943-KK/FF, January 1972.
16. KAGAN, PAUL ASSOCIATES, INC. *Cablecast*. No. 86 (June 5, 1973).
17. ———. *Pay TV Newsletter*. Rockville Centre, N. Y.: July 24, 1973.
18. MITCHELL, B. M. AND SMILEY, R. H. "Cable Television under the 1972 FCC Rules and the Impact of Alternative Copyright Fee Proposals: An Economic Analysis." Mimeograph, National Cable Television Association, Washington, D C., September 20, 1972.
19. MYERS, S. C. "The Application of Finance Theory to Public Utility Rate Cases." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 58–97.
20. NOLL, R. G., PECK, M. J., AND MCGOWAN, J. J. *Economic Aspects of Television Regulation* Washington, D. C.: The Brookings Institution, 1973.
21. PARK, R. E. "The Exclusivity Provisions of the Federal Communications Commission's Cable Television Regulations." The Rand Corporation, R-1057-FF/MF, June 1972. (Summarized in "After Exclusivity Blackouts, What's Left on the Horizon," *TV Communications*, August 1972.)
22. ———. "Potential Impact of Cable Growth on Television Broadcasting." The Rand Corporation, R-587-FF, October 1970. (Summarized in "The Growth of Cable TV and Its Probable Impact on Over-the-Air Broadcasting," *The American Economic Review*, Vol. 61, No. 2, May 1971.)
23. ———. "Prospects for Cable in the 100 Largest Television Markets." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 130–150.
24. RIVKIN, S. R. *Cable Television: A Guide to Federal Regulations*. The Rand Corporation, R-1138-NSF, March 1973
25. SEIDEN, M. H., AND ASSOCIATES. *CATV Report*. Washington, D. C., 1970.
26. SHARPE, W. F. *Portfolio Theory and Capital Markets*. New York: McGraw-Hill Book Company, 1970.
27. SLOAN COMMISSION ON CABLE COMMUNICATIONS. *On the Cable: The Television of Abundance*. New York: McGraw-Hill, 1971.
28. TELEVISION DIGEST, INC. *CATV and Station Coverage Atlas*. Washington, D. C., 1971–1972 ed.
29. ———. *Television Factbook, Services Volume*. Washington, D C., 1970–1972.
30. U. S. FEDERAL COMMUNICATIONS COMMISSION. "Cable Television Service: Cable Television Relay Service." *Federal Register*, Vol. 37, No. 30, Part II (February 12, 1972), pp. 3252–3341.
31. ———. "Cable Television Service: Reconsideration of Report and

Order." *Federal Register*, Vol. 37, No. 136, Part II (July 14, 1972), pp. 13848–13910.

32 ———. "Second Further Notice of Proposed Rulemaking." *FCC Reports* (August 21, 1970), pp. 580–589.

33. WEINBERG, G. *Cost Analysis of CATV Components: Final Report.* RMC Report UR-170, June 1972, prepared for the Office of Telecommunications Policy.

A reexamination of the prophecy of doom for cable television

Robert W. Crandall

Associate Professor of Economics
Massachusetts Institute of Technology

and

Lionel L. Fray

Senior Executive Consultant
Temple, Barker & Sloane, Incorporated
Wellesley Hills, Massachusetts

The economic literature on the prospects for cable television in the nation's major television markets is universally pessimistic. The conventional wisdom, as expressed by Park and Mitchell in the pages of this Journal, is that cable systems will not be able to enroll many subscribers in areas served by three commercial stations. Given that the FCC has limited the importation of distant independent signals in a number of ways, the prospective additions to a viewer's television menu from cable television are thought to be quite limited in areas of three or more stations. This traditional view is challenged in this paper, for it is the authors' opinion that the potential of cable television is greatly underestimated if attention is focused solely upon importations of broadcast signals and low-quality local originations. Cable television opens up the possibility for a price-rationed system of entertainment in the home which far surpasses the current television medium. When reasonable assumptions, drawn from the cable industry and the capital market, are inserted into the Comanor-Mitchell model of cable-system performance, rates of return on prospective investment are consistently above the cost of capital. One needs only to assume that pay cable will allow systems to attract 50 percent of the homes passed by the cable and that revenues per subscriber will increase at a rate of 2 percent per year in order to obtain this more sanguine result. Such assumptions seem far more realistic as the basis for long-run projections of the future of cable television.

1. Introduction

■ In the past four years, the cable television industry has witnessed continuing growth in subscriber levels. The number of primary subscribers has risen from approximately 4.5 million at the beginning of 1970 to over 7 million at present. Moreover, the FCC's Cable Television Report and Order of February 1972 has

---

The research reported in this paper was financed by the Motion Picture Association of America, but the conclusions reached by the authors are their own and do not necessarily reflect the opinions of the M.P.A.A. or its individual member companies.

opened up the largest 100 markets to distant signal importation and has signaled a rapid rise in cable activity in these markets. The average number of franchise applications per central city in the top 100 markets has nearly doubled since 1970, and no fewer than 22 systems began construction in these central cities in the 1970 to 1972 period.[1] Concurrently, there has been an active market for acquisitions of existing operative systems—a market in which transfer prices have averaged more than 150 percent of reproduction costs.[2]

These trends are rather unremarkable given the current size as contrasted with the enormous potential of cable television. What makes them notable, however, is the sharp contrast which they provide to the forecasts which have appeared on the pages of this *Journal.* Comanor and Mitchell predicted in early 1970 that few cable systems would earn as much as 12 percent on capital before taxes in the event the FCC adopted its proposed rules on signal importations, capacity requirements, commercial substitution, public broadcasting fees, and copyright fees.[3] Even gloomier forecasts were to follow. One year later, Park[4] published a study of the demand for cable services which predicted subscriber penetration levels far below those determined by Comanor-Mitchell (hereafter, C-M). Mitchell,[5] utilizing this study, and revising the earlier C-M simulations to incorporate the 1972 FCC Rules,[6] reaches the conclusion that only the very largest systems could hope to earn as much as 12 percent before taxes in the largest 100 markets even if copyright fees were not imposed upon cable operators for distant-signal importations. Given that the cost of capital to these firms is higher than 12 percent before taxes, not only would cable operators be unable to pay copyright fees, but—according to these forecasts—they could look forward to imminent collapse in the face of the cruel economics of this unattractive business.

Clearly, these economic forecasts must be in error or there are numerous investors courting bankruptcy by rushing into the quagmire of large-market cable services. While we cannot claim to be sufficiently prescient to make precise estimates of the rates of return to be expected on future investment in cable television, in this paper we shall demonstrate that the pessimistic conclusions of C-M and Mitchell are dictated by the choice of equally pessimistic forecasts of two critical parameters: (1) the number of households which will enroll as cable subscribers in the nation's largest markets and (2) the prospective rate of growth of revenues per cable subscriber. We shall replace these forecasts with a range of values drawn principally from the cable industry itself and thereby generate a range of potential rates of return. Less important sources of downward bias due to overstatement of operating expenses—especially for the smaller systems—and to projections of excessive underground construction will also be eliminated. With these changes, we are able to show that the C-M and Mitchell prophecy—the death of cable television in the largest television markets from the cancer of low profitability—is open to serious question.

---

[1] See [11].

[2] Current sales of cable franchises are occurring at prices of approximately $400 per subscriber for fairly mature systems. An econometric analysis of sales involving approximately one-tenth of all subscribers in 1971–1972 reveals that the average price paid per subscriber at maturity is more than $320. (See Crandall [4].) According to industry estimates, reproduction costs of the assets purchased are no more than $200 per subscriber.

[3] See Comanor and Mitchell [3].

[4] See Park [9].

[5] See Mitchell and Smiley [6].

[6] See U. S. FCC [12].

2. The Comanor-Mitchell and Mitchell estimates

■ It is useful to begin with a brief review of C-M and Mitchell. Both studies seek to estimate the rate of return on cable investment assuming perpetual franchise life and periodic replacement of all physical assets. The Mitchell study uses the cost assumptions of C-M, modified to take account of the actual cable rules promulgated by the FCC after C-M was published. Assumptions concerning demand and revenues per subscriber are somewhat different in the two studies.

Table 1 summarizes the important assumptions employed in each paper. The precise assumptions employed in both studies concerning operating costs and capital costs per mile of cable underground and above ground may be found in Appendix 3 to C-M. This table is rather self-explanatory except for the demand equations employed. C-M estimated a log-linear demand function from a sample of 149 cable systems drawn from the *Television Factbook*. While the estimated equation which they used to predict future subscriber demand has very poor statistical properties and several coefficients with implausible signs, it provided reasonable estimates of ultimate "mature" subscriber penetration.

TABLE 1

ASSUMPTIONS EMPLOYED BY COMANOR-MITCHELL AND MITCHELL

	COMANOR-MITCHELL	MITCHELL
1. DEMAND FUNCTION	LOG-LINEAR	PARK'S LOGIT FUNCTION
2 RANGE OF SUBSCRIBER PENETRATION ESTIMATES AT MATURITY	TOP 100 59% OUTSIDE TOP 100 63%	TOP 100 24.0% TO 44 9% OUTSIDE TOP 100 29 5% TO 65 6%
3 REVENUES PER SUBSCRIBER AT OUTSET	$64 60 or $69 00	$64 60
4 RATE OF GROWTH OF REVENUES PER SUBSCRIBER	ZERO	ZERO
5 HOMES PER MILE OF CABLE	80-120	80-200
6. PERCENTAGE OF CABLE UNDERGROUND	5-20	0-20
7 LIFE OF PLANT	10-15 YEARS	15 YEARS

In his revision of the original C-M profitability estimates, Mitchell uses the demand function estimated by Park instead of the original C-M function. While Park's estimated demand equation has far more satisfactory statistical properties, it is subject to similar theoretical difficulties which we detail below. Moreover, it projects subscriber penetration rates which are implausibly low given the expectations of investors and system operators.[7] Moreover, Mitchell uses the more conservative estimate of \$64.60 in revenues per subscriber per year *ad infinitum*, dropping the \$69.00 version from C-M. Thus, Mitchell's demand forecasts are much more conservative than those derived from C-M, and these more than offset the more realistic population density assumptions which he employs—80 to 200 homes per mile of trunk cable as compared with 80 to 120 for C-M.

C-M estimated the rates of return in prospect for cable firms which satisfied the proposed FCC Rules (the so-called "Public Dividend Plan"), enunciated in 1969 and 1970. These rules would have required cable companies to:

(1) provide twenty-channel capacity;
(2) be able to offer two-way capability;
(3) provide local originations if 3,000 or more subscribers are enrolled;
(4) substitute local-station commercials for commercials on imported distant signals;
(5) pay 5 percent of gross revenues to a fund for public broadcasting if the system is larger than 3,500 subscribers and in the top 100 markets;
(6) pay annual fees to the FCC and comply with technical-standard certification requirements; and
(7) pay copyright fees, in a manner to be determined by future legislation.

In return, the FCC offered systems in the top 100 markets the opportunity to import distant signals, a privilege which had been denied pending the FCC's examination of the effect of cable television upon off-the-air television broadcasting.

In 1972, the FCC announced its Cable Television Rules, which incorporated a now famous compromise on signal importation, exclusivity protection, copyright fees, and cable originations. Briefly, the Commission now allows cable systems in the top 100 markets to fill out their offerings to include three network stations and either three independents (in the top 50 markets), two independents (in the second 50 markets), or one independent (in markets 101+). In all of the top 100 markets, local stations' purchases of syndicated programming will be protected for as much as 2 years by the requirement that certain identical programs on imported signals be blacked out. This "exclusivity protection" clearly reduces the attractiveness of imported signals because it forces cable companies to offer fewer than one to three independents during the time the imported independents are offering syndicated programming subject to the exclusivity protection or, alternatively, it requires the cable system to import "stand-by" signals to fill in channels blacked out by this provision of the rules. The FCC also followed through on its intention to require some originations of all systems of 3,500 or more subscribers. The 1972 Cable Rules delete the commercial substitution and 5-percent public broadcasting fee requirements but maintain the twenty-channel capacity requirement for the top 100 markets and the two-way (latent) capability requirement for all markets. Finally, it was left to the contesting parties and to the Congress to work out an appropriate copyright-fee schedule.

---

[7] See Crandall [4] for estimates of the value attaching to unbuilt portions of large systems in major cities.

In neither the Public Dividend Plan nor the final Rules did the FCC settle upon a schedule of copyright fees to be paid by cable operators to suppliers of off-the-air programming. In C-M, a set of proposed fees advanced in a current Senate Bill, requiring fees of 1 percent to 5 percent of gross revenues, is included in the costs of meeting the proposed FCC regulations. In Mitchell's estimates, various copyright fee schedules—including this 1- to 5-percent variant—are included in separate calculations.

The rates of return on investment before corporate income taxes which C-M and Mitchell project for the proposed and actual FCC Rules, respectively, are reproduced in Table 2. There can be little doubt that the C-M projections are consistently below the most conservative estimate of the cost of capital before taxes. Mitchell's calculated rates of return are only slightly more attractive. With a 1- to 5-percent copyright fee schedule, there would be no systems in the top 100 markets earning so much as 12.5 percent before taxes; with no copyright fees at all, only 50,000 subscriber systems would earn as much as 15 percent on capital before taxes. Therefore, we must presume that it is the combined prediction of these studies that the FCC Rules have sounded the death knell to cable television in the nation's largest 100 markets, where over two-thirds of the nation's homes are located. The subsequent imposition of copyright fees could only serve to destroy the prospects for those few systems able to survive for some atypical reason despite the FCC Rules.

3. A review of the critical assumptions

■ **Demand.** Both C-M and Mitchell project future subscriber levels by employing demand functions which have been estimated econometrically across a sample of systems providing virtually no services other than the retransmission of broadcast signals. It is highly likely, however, that subscriber demand in the largest markets will be influenced significantly by additional services in the form of program originations such as recent motion pictures and sports events not currently available on commercial television. For this reason alone, these econometric studies cannot provide a sound basis for projecting further subscriber demand. But as we shall demonstrate, even if the environment were not to change, there are serious difficulties in the use of their results.

C-M utilize a large sample of CATV systems drawn from the *Television Factbook* to estimate a log-linear demand curve by ordinary least squares. Park's study, which Mitchell utilizes in

TABLE 2

INTERNAL RATES OF RETURN PROJECTED BY COMANOR-MITCHELL AND MITCHELL FOR VARIOUS MARKET CONDITIONS

COMANOR-MITCHELL* (PROPOSED RULES, INCLUDING 1-5% COPYRIGHT FEES)			MITCHELL (ASSUMING 1-5% COPYRIGHT FEES AND 1972 CABLE RULES)		
MARKET	SYSTEM SIZE AT MATURITY (SUBSCRIBERS)	RATE OF RETURN (%)	MARKET	SYSTEM SIZE AT MATURITY (SUBSCRIBERS)	RATE OF RETURN (%)
TOP 100 MIDDLE	7,500 10,000 25,000 50,000	3 85 5 15 9 16 11 25	TOP 50 MIDDLE	10,000 25,000	6 5-9 3 3 4-5 4
TOP 100 EDGE	5,000 7,500 10,000 25,000 50,000	0.47 4 83 6 33 10 95 13 40	TOP 50 EDGE	10,000 25,000	5 7-6 6 9 9-11 4
MARKETS 101+	5,000 10,000 15,000	5 78 11 19 13 31	51-100 MIDDLE	10,000 25,000	1 7-5.4 5 0-9 1
			51-100 EDGE	10,000 25,000	4 8-7 3 8 6-12 2
			MARKETS 101 MIDDLE	7,500** 15,000	5 9-15 4 8 2-18 6
			MARKETS 101-EDGE	7,500** 15,000	7 1-14 6 9 7-17.8

*ALL CALCULATIONS ASSUME·
1 15-YEAR LIFE
2 STANDARD ORIGINATION
3 POLE RENTAL OF $175 to $250 PER MILE PER YEAR
4 MAXIMUM C-M ADVERTISING REVENUES OF $6.60 PER SUBSCRIBER PER YEAR FOR MARKETS 101+, $2 20 PER SUBSCRIBER PER YEAR FOR TOP 100 MARKETS
5. MAXIMUM SIGNAL IMPORTATIONS
**MINIMUM ORIGINATION AND SUBSCRIBER REVENUES OF ONLY $62 40 PER YEAR.

his revision of the profitability estimates, uses a much smaller sample of systems (63) located in areas of good reception, similar to the quality of reception to be expected in the center of most of the nation's top 100 markets. His sample data are obtained by direct response from cable owners, not from the *Factbook*, but his method is also to utilize ordinary least squares in estimating a nonlinear demand function. While the two estimation procedures are different—Park is forced to estimate his nonlinear equation by a trial-and-error method of successive approximation—both studies suffer from at least one common defect: they assume all independent variables to be predetermined and not a function themselves of the expected subscriber penetration of the system. Since cable franchises are awarded by local municipalities by selecting from a number of different bids in most instances, it is plausible that price will be set on the basis of expected subscriber penetration. Higher penetration for given cable plant investment lowers the cost of performing the cable services, and, therefore, the price which cable owners must charge in order to cover their full costs. Estimating demand functions by assuming that price is predetermined will lead to biased estimates of the price coefficient. In fact, Park's discovery that there exists a unitary elastic relationship between subscriber penetration and price is consistent with the theory that the marginal cost of connecting another subscriber to an existing cable plant is very nearly zero, that municipalities are efficient in their regulation of local cable franchises, and that subscriber demand is perfectly inelastic. In such a set of circumstances, for a given service configuration and plant investment, a municipality would require that total revenues from subscriber fees be fixed at a constant level, thus leaving the impression in a single-equation model of "unitary elastic demand."[8]

An equally serious simultaneity problem occurs in including the number of homes passed as a deflator for subscribers in the measure of "subscriber penetration" in the dependent variables utilized in both studies. The number of homes passed typically depends upon the age of the system since no system can be wired overnight. This is especially serious in Park's study which has systems varying in size from 227 homes to 20,000 homes. Estimating the temporal path of subscriber enrollment in each study requires estimating a relationship such as:

$$S = [f(P,Y,X,\ldots.)g(t)]H, \qquad (1)$$

where S is the number of subscribers, P is the subscription price, Y is the income per household, t is calendar time from the date the system was completed, and H is the number of homes passed by the cable plant. Estimation is actually effected only after dividing both sides of (1) by H:

$$\frac{S}{H} = f(P,Y,X,\ldots.)g(t), \qquad (2)$$

and assuming H to be independent of t. If the plant were fully built *ab initio*, this approach would be appropriate, ignoring the bias from simply regressing S on P. But it is unreasonable to assume that H is fixed from the very start; therefore, estimating (2) may lead to a bias in the coefficient of t if immature systems are included in the sample. This is especially serious for the Park study if a number of systems have spread their cable construction over more than, say, one year, as a simple example will demonstrate.

Assume that a given cable system chooses to spread its investment out over 3 years in five discrete segments. The first segment

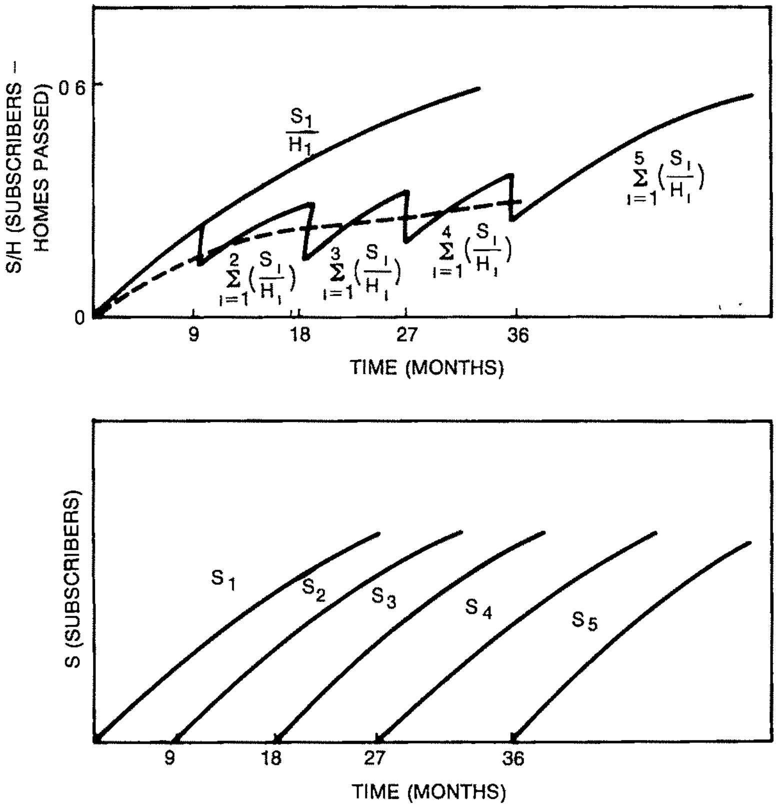
---

[8] It may be that ordinary least squares estimates of the relationship between subscriber penetration and the monthly fee are appropriate for forming predictions of future penetration and that structural demand estimates are not required, but this requires one to assume that the regulatory climate will not change—i.e., that the relationship between prospective penetration and the fees which municipal authorities will allow in the future is the same as in the past. Moreover, we do not contend that the resulting unitary elastic demand equation is an unreasonable representation of reality. We merely contend that all extant estimates of price elasticity of *demand* based upon single-equation estimation techniques are vulnerable to an identification problem.

is completed at month zero, the second is completed at month 9, and so on through month 36. In addition, assume that subscribers in each area are only enrolled beginning with that segment's completion date. In each case, assume there exists a nonlinear path $g(t)$, which explains the approach of subscriber enrollment to eventual maturity. Each of these hypothetical paths is plotted as $S_1, S_2, \ldots, S_5$ in the bottom panel of Figure 1. If the extra system

FIGURE 1

PENETRATION PATH



comprised only segment one, the path of penetration would resemble S_1/H_1 in the top panel of Figure 1. On the other hand, if the system is built in the five hypothesized segments, we would observe a jagged approach of penetration—defined as subscribers per home passed *currently* by the cable—toward maturity. If a sample of systems was comprised of a large number of immature systems of this type, estimating a linear or log-linear maturation path from such a pattern (as the Comanor-Mitchell and Park studies do) would yield a distorted picture of reality. In particular, there would seem to be little growth in penetration from a sample of points drawn at random from the jagged path from months 9 through 36. Thus, one might conclude that little further growth is expected from such systems, and that the dashed "maturity path" is the true path of subscriber enrollments. In addition, one would expect the relationship to be strongly heteroscedastic with greater variance in the months 9 through 36 than in later months.

When Park estimates his equation, he finds that the best-fitting "maturation" path is linear through month 18 at which time the systems are "mature." Could his results have been dictated by a protracted pattern of construction such as we hypothesize? We cannot answer this question because Park does not provide evidence on the pattern of the residuals from his equation. Since his data were obtained on a confidential basis we cannot replicate his results, but we can draw a random sample from the most recent *Factbook* to test the predictive power of his and C-M's estimated equations.

Our sample of CATV systems is obtained by selecting a system at random from the listings by state in the 1972–1973 *Factbook* and selecting every twentieth system thereafter. This procedure yields a total sample of 153, but only 81 of these reported sufficiently complete data to permit fitting the alternate demand equations: Park's equation with linear growth through 18 months; Park's equation as modified by Mitchell to allow for nonlinear growth over 5 years; and C-M's equation. None of the demand equations proves to be a very good predictor of penetration. Table 3 provides the appropriate data on bias and goodness-of-fit for all 81 systems and the 20 which are located in the top 100 markets. There is a downward bias evident in each equation, sometimes exceeding one-third of actual penetration levels. Moreover, the root mean square errors are very large, suggesting very low predictive power of all three equations.[9] In light of this evidence, we feel justified in asserting that none of the equations should be utilized in future projections of profitability.

TABLE 3

THE PERFORMANCE OF ALTERNATIVE STATISTICAL DEMAND FUNCTIONS

	PARK-LINEAR GROWTH	PARK-MITCHELL'S GROWTH	COMANOR-MITCHELL	ACTUAL
ALL MARKETS				
MEAN PENETRATION	0.498	0.516	0.402	0.560
DOWNWARD BIAS	0.062	0.044	0.158	—
STANDARD DEVIATION	—	—	—	0.223
ROOT MEAN SQUARE ERROR	0.260	0.290	0.270	—
TOP 100 MARKETS				
MEAN PENETRATION	0.341	0.322	0.331	0.448
DOWNWARD BIAS	0.106	0.126	0.117	—
STANDARD DEVIATION	—	—	—	0.182
ROOT MEAN SQUARE ERROR	0.249	0.299	0.262	—

---

[9] To be fair to Park, we must admit that his results were based upon a sample of systems in areas with little or no reception problems. Our sample undoubtedly contains a number of systems in areas of poor signal quality One must ask, however, how many households in the top 100 markets will be faced with high quality reception without outside antennas. The construction of tall buildings in the center of these markets has already created severe reception problems for a number of viewers, and increasing reliance upon high-rise structures promises to make local reception even more problematical.

Given that none of the extant studies allows for the potential of new cable services, it is even more likely that any projections based upon them would be too conservative even if they were properly estimated. A plausible substitute for these estimates is a set of projections derived from large multiple-system owners and published by the National Cable Television Association—which also sponsored the C-M and Mitchell studies. These estimates place subscriber penetration in the top 100 markets at 51 percent of homes passed by 1982, 59.7 percent by 1987, and 67.3 percent by 1992.[10] If we are interested in making profitability projections for a plant of 15-year life, we might choose the *mature* penetration levels implied by these data for 1987—65 percent —and the average for current systems—50 percent—as a reasonable range for prospective subscriber penetration in the top 100 markets. As existing systems add new services, import new signals permitted by the FCC Rules, and simply mature, it is not unreasonable to project that penetration rates will increase and that the N.C.T.A. projections will be vindicated.

☐ **Revenues per subscriber.** Both C-M and Mitchell have a myopic view of the potential revenue per subscriber available to cable owners during the infinite franchise life utilized in their calculations. While the monthly fees may not grow very rapidly from the current average of $5.25 per month for a primary connection, even these fees will probably grow at a rate exceeding zero percent. Both C-M and Mitchell allow for some cost inflation through escalation of salaries, but they fail to allow for price increases which are not infrequent in the industry. Our 81-system sample, reported above, demonstrated an annual rate of growth in the average monthly subscriber fee of 1.2 percent per annum for the past 5 years. Given a lagged adjustment to recent inflation inherent in any quasi-regulatory scheme, it is not unreasonable to assume that this growth rate may rise in the near future.

More important is the growth in revenue per subscriber which is likely to come from ancillary sources such as motion-picture channels, special cultural events, sports packages, and a number of other offerings. While it is presently impossible to muster precise data on the prospective yield from these services, inferences may be drawn from the relationship between prices paid for cable system acquisitions and the reproduction costs of the assets required.

An econometric investigation into the purchase price of assets acquired in cable company mergers and sales in 1971–1972 reveals that buyers were paying slightly more than $320 per subscriber at maturity.[11] With evidence on current cash flow and the opportunity cost of capital, we may deduce the rate of growth in cash flow anticipated by buyers and sellers of these existing assets from the following relationship:

$$\frac{PV}{MS} = \int_0^\infty [(R_0 - OC_0 - I_0)e^{gt}]e^{-rt}\,dt, \tag{3}$$

---

[10] See Andersson [1].

[11] Crandall [4] is the source for the discussion which follows.

where PV is present value or sales price, MS is the number of subscribers projected for the system at maturity, R_0 is current revenues per subscriber, OC_0 is current operating costs per subscriber, I_0 is the current annual cost of replacing assets as they depreciate, g is the anticipated perpetual rate of growth of cash flow, and r is the opportunity cost of capital before corporate income taxes. The expression (3) may be integrated and evaluated over an infinite horizon at:

$$\frac{PV}{MS} = (R_0 - OC_0 - I_0)/(r - g). \tag{4}$$

Operating costs for current systems which are fairly mature are equal to 0.50 to 0.60 of gross revenues.[12] For a mature system, we may presume that this ratio is no higher than 0.50 since operating costs do not rise proportionally with subscribers for a given plant size. Investment costs for mature systems are more difficult to estimate, for we cannot know whether the systems in the sample generating a \$320 present value per mature subscriber are refurbished continuously or only at the end of discrete periods or "generations." C-M assume and current accounting practices generally require a 15-year useful life for the capital equipment invested in cable systems. Since these assets total \$200 per current subscriber, it is reasonable to assume that they are no more than \$180 per subscriber at maturity. If 1/15 of these assets must be replaced each year, the annual reinvestment costs are \$12 per subscriber. If the reinvestment occurs only at the end of each 15-year period, the annual capital charge is a function of the assumed cost of capital. The final calculation at a discount rate of 0.15 to 0.16 before corporate taxes for the annual charge required to refurbish the assets of a system one-half generation old in 7½ years and every 15 years thereafter is between 0.053 and 0.055 of \$180, or slightly less than \$10. With current annual revenues per subscriber averaging \$63 per first connection plus additional revenues from installations and secondary connections, investment costs may be estimated at one-sixth of current revenues. Thus, current cash flow is equal to 0.33 of current revenues, or \$21 to \$23, depending upon the importance of installation fees and secondary connections. Therefore, we may estimate $(r - g)$ as lying in the range of \$21/\$320 to \$23/\$320, or 0.066 to 0.072.

The cost of capital before taxes is equal to the weighted average of equity capital and debt capital. In 1971 to 1972, the largest cable firms had a debt-equity ratio of approximately 1:2 and were facing borrowing rates of at least 0.085 for long-term debt capital. Since the five largest firms had betas (from the capital-asset pricing model) of 1.8,[13] we estimate that the after-tax equity cost was 16.8 percent and the before-tax return demanded by holders of their common equities was 19.6 percent.[14] Thus, the cost of capital for a firm of this leverage may be calculated at 15.8 percent and the anticipated rate of growth of cash flow is approximately 9 percent per annum *ad infinitum.*

---

[12] See the next section for a discussion of operating costs.

[13] See Myers [7] for a discussion of the application of the capital-asset pricing model to the estimation of the cost of capital.

[14] If ρ_{at} is the rate of return demanded by holders of equity after corporation income taxes, ρ_{bt}—the rate of discount before taxes—must be equal to

The projected increase in cash flow from the sale of leased channels or the offering of special entertainment packages or other services will undoubtedly be the result of large increases in both revenues and costs. Converters, special headend equipment, additional billing expenses, and additional costs of interconnection will require considerable outlays not anticipated in current cost analyses, but it is obvious that investors expect revenues to outpace these additional costs by a wide margin. Since we cannot possibly forecast the precise rate of growth of both costs and revenues, in the analysis below we shall allow revenues to grow at a modest rate but assume an unchanging cost function. At a 15.8-percent discount rate, a 9-percent rate of growth in cash flow for 20 years would be approximately consistent with constant costs per subscriber at maturity and a growth in revenues of 3 to 4 percent per annum.

Total current revenues per subscriber—excluding installation fees—are equal to $63 per year plus an unknown additional contribution for secondary connections. Since 41 percent of television homes currently have second sets, we assume that 41 percent of cable subscribers will pay $12 per year for an additional connection. Total revenues are thus taken to be $67.92 per subscriber at present. In the calculations below, we shall assume these revenues grow at rates of zero, 2 percent, and 4 percent per annum reflecting the contribution of ancillary services and fee increases. Implicitly, therefore, we—like Mitchell—assume a zero growth rate for net revenues per subscriber from the basic cable service of retransmission of broadcast signals. We cannot, however, ignore the considerable potential of new services by simply adding on $2.20 or $6.60 per year in advertising revenues per subscriber as Mitchell and C-M have done. Instead, we ignore advertising as a transitory influence which may eventually be dominated by cable offerings of either leased channels to independent "pay television" networks or individual subscription channels offered by the systems directly to subscribing households. Either variant promises much higher prospective revenues than $2.20 or $6.60 per subscriber per year, a fact which has not escaped the notice of recent buyers of cable systems.

$$\frac{\rho_{at}\left[\frac{\Pi}{V}+g'\right]}{\left[\frac{\Pi\,(1-\Upsilon)}{V}+g'\right]},$$

where Π/V is the ratio of net profits before taxes to the value of equity, Υ is the marginal rate of corporate taxation, and g' is the rate of growth of profits. Π/V is currently equal to about 0.055 for a firm with $214 of equity per mature subscriber, and $\Pi(1-\Upsilon)/V$ is equal to approximately 0.028 under these assumptions, reflecting a P/E ratio of nearly 40:1. Thus, g' must be equal to 0.14 and $\rho_{bt}=0.196$.

☐ **Operating costs.** The annual operating costs of cable systems derive from the requirements of maintenance, connecting and disconnecting subscribers, operation of the headend, billing of customers, and production of locally originated programming. The least of these functions had been relatively unimportant until the recent FCC Rules were promulgated, requiring some local originations for all systems above 3,500 subscribers. Previously, most systems offered limited live-tape originations, confining themselves to automated time-news-weather formats.

There are very few studies of the determinants of CATV operating costs. By far the most extensive research into cable costs is that undertaken by C-M.[15] Since that time, a number of other studies have appeared, but virtually all of these rely upon the C-M parameters to some degree.[16] The C-M estimates are based upon a variety of *a priori* engineering assumptions concerning input requirements for systems of various sizes, but there has been no statistical analysis of these implicit assumptions about the production function for cable nor of the cost functions which derive from it. Since we shall be forced to use a variant of the C-M cost data in estimating system costs below, we first pause to examine the validity of their estimates. For this purpose, we compare their findings with actual operating results reported by cable companies whose securities are traded with the public.

Most cable companies are so small that their securities are not publicly traded and their income statements are not reported. A few large multiple system owners (MSO's) are listed on the major securities exchanges, therefore, their operating results are published periodically. Of these MSO's, however, several are so diversified as to make it impossible to separate cable costs from the costs of other operations. Among these firms are TelePrompTer, Viacom, and Warner Communications. In searching listing statements, prospectuses, and other financial services, we found five virtually undiversified MSO's which reported operating costs for all or part of the period 1969 to 1972. Of these firms, two—H&B American and Cypress—have subsequently been acquired by other MSO's and one—Cox—has recently been forced to abandon its merger with A.T.C. The others, CableCom General and Columbia Cable Systems, continue as independent entities.

The five MSO's which we analyze had a combined total of more than 770,000 subscribers at the end of the reporting periods cited.[17] Our first task, therefore, is to construct measures of average system size and operating costs per subscriber. Since average penetration—subscribers enrolled as a proportion of homes passed by the cable—for these companies ranged from 0.492 to 0.642 (see Table 4), we may safely surmise that the systems owned by these companies were quite mature. This is an important consideration because start-up costs and the accounting of them vary considerably across systems, and it would be difficult to correct for these differences.

The calculation of average system size is based upon end-of-

---

[15] In [3].

[16] See, for example, Baer and Park [2] and Weinberg [13]

[17] See Table 4.

TABLE 4

OPERATING COSTS FOR FIVE MSO's 1969 TO 1972

COMPANY	PERIOD	AVERAGE PENETRATION (SUBSCRIBERS ÷ HOMES PASSED)	AVERAGE HOMES PER MILE	TOTAL SUBSCRIBERS	AVERAGE SUBSCRIBER SIZE	COST PER SUBSCRIBER
A. CABLE COM GENERAL	9/69 TO 6/70	0.499	75	103,000	3,800	$37.20
B. COX	1970	0.506	95	197,000	14,900	36.56
C. H&B AMERICAN	7/69 TO 12/69	0.537	86	245,000	6,300	34.90
D. CYPRESS	6/71 TO 3/72	0.642	83	165,000	11,400	33.93
E. COLUMBIA CABLE SYSTEMS	9/71 TO 3/72	0.492	81	71,023	8,500	37.04

SOURCES

A AMERICAN STOCK EXCHANGE LISTING APPLICATION NO 8797, OCTOBER 16, 1970

B AMERICAN STOCK EXCHANGE LISTING APPLICATION NO 9031, MARCH 18, 1971

C AMERICAN STOCK EXCHANGE LISTING APPLICATION NO 8739 (TELEPROMPTER), SEPTEMBER 14, 1970

D NEW YORK STOCK EXCHANGE LISTING APPLICATION NO B-2188, SEPTEMBER 12, 1972

E. SECURITIES AND EXCHANGE COMMISSION, REGISTRATION STATEMENT, S-1, NO 2-44751, JUNE 23, 1972

TELEVISION FACTBOOK, ANNUAL ISSUES

period subscribers reported by these firms in their prospectuses, or, in one case (Cox), in the *Television Factbook*. The "average" size of system reported in Table 4 is a weighted average, using the number of subscribers as weights.

Operating costs per subscriber cannot be measured by dividing end-of-period subscribers into reported operating costs because new subscribers are added during the year, some by expansion of given systems and others through mergers and acquisitions. For this reason, an average subscriber level is calculated first by dividing the weighted average annual revenues per subscriber—equal to 1.06 times 12 months' subscriber fees[18]—into reported revenues from operations. This subscriber level is then divided into total operating, administrative, and selling costs for the most recent period reported. Depreciation, amortization, and interest are excluded since they reflect capital costs. The resulting estimates of operating costs per subscriber and average system size are reported in Table 4.

---

[18] Six percent is added to subscriber fees reported in the *Factbook* to account for installation fees and secondary connections in 1970. Installation fees average approximately $10, but for mature systems they are collected from less than 20 percent of subscribers in each year. Data on secondary connections are not available, but it is assumed that 25 percent of subscribers in 1970 paid $12 per year for secondary connections. Thus, secondary connections are assumed to add an average of $3 per subscriber per year. Since installation fees are frequently rebated and even one month's subscriber fees are waived for new connections, only $1 per subscriber is added for the net contribution of installation fees. Given average monthly fees of $5.25, these additional revenues of 33.3¢ per month add 6 percent to calculated subscriber fees.

The estimates of operating costs for the four MSO's demonstrate remarkably little variance given the wide range of average system size—from 3,800 to 14,900 subscribers. Unfortunately, direct comparison is impossible, given the differences in average subscriber penetration and population density across the companies, but it appears unlikely that there are important economies of scale in these firms' cable operations. In order to compare these estimates with C-M predictions, we have projected estimates for each—given average size, penetration, and population density—assuming that each imported three signals by microwave, provided standard automated originations, and paid pole and tower rentals of $250 per mile and franchise fees of 5 percent of total revenues. These estimates are shown in Table 5 beside the financial statement calculations, measured in 1970 dollars assuming a 5-percent annual rate of cost inflation from 1969 through 1972. The projected errors from the Comanor-Mitchell procedure show a striking inverse correlation with system size. In Figure 2, the scatter of these errors against average subscriber level demonstrates that Comanor and Mitchell appear to overestimate system costs for very small systems and to underestimate costs for very large ones. This is not an uncommon result for engineering estimates of production functions, for this methodology often exaggerates the economies of scale inherent in economic processes. In this case, it appears that C-M methodology for pre-1972 rules is most accurate for systems of about 10,000 subscribers, but clearly in error for systems of fewer than 5,000 subscribers. The errors would be somewhat less striking if we were to drop the assumption of signal importation by microwave, but most systems imported some signals by microwave in the 1969 to 1972 period; some imported many more than three distant signals in this manner.

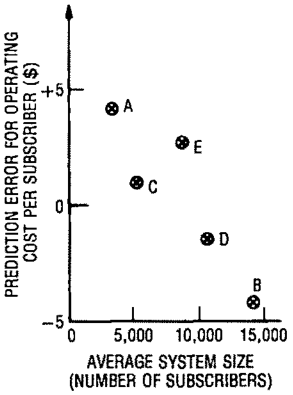
TABLE 5

ACTUAL AND COMANOR-MITCHELL PREDICTED VALUES FOR OPERATING COSTS FOR FIVE MAJOR MSO's

COMPANY	OPERATING COSTS PER SUBSCRIBER		COMANOR-MITCHELL PROJECTIONS	PROJECTED ERROR IN 1970 DOLLARS
	CURRENT DOLLARS	1970 DOLLARS (ASSUMING 5% ANNUAL INFLATION)		
A CABLECOM GENERAL	$37 20	$37 90	$44 48	$6 58
B COX	36 56	36.56	32 49	−4 07
C H&B AMERICAN	34 90	36.14	37 11	0.97
D CYPRESS	33 93	31 75	32 39	−0 64
E COLUMBIA CABLE SYSTEMS	37 04	34 26	37 35	3 09

FIGURE 2

PREDICTED ERROR VERSUS AVERAGE SYSTEM SIZE FROM COMANOR-MITCHELL COST DATA



A further exercise may be employed to demonstrate the scale economies inherent in the C-M parameters. In Table 6, we have calculated the operating costs for various sizes of system assuming an average density of 90 homes per mile and 50-percent pene-

TABLE 6

OPERATING COSTS — COMANOR-MITCHELL METHOD

SYSTEM SIZE AT MATURITY	COMANOR-MITCHELL RESULTS[a]	OPERATING COSTS PER SUBSCRIBER	
		WITHOUT 1972 FCC RULES[b]	WITH 1972 FCC RULES[c]
25,000	—	$29 45	$31 94
15,000	$26 72	32 97	36 94
10,000	27 16	34 91	40 72
5,000	31 49	41 01	51 38
3,500	34 97	46 15	62 21

[a] FROM COMANOR-MITCHELL, OP CIT , APPENDIX J, "CURRENT REGULATIONS," WHICH ASSUME NO ORIGINATIONS OR DISTANT SIGNAL IMPORTATION

[b] ASSUMES

1 MATURE PENETRATION = 0 50
2 MATURITY ACHIEVED IN YEAR 5
3 AUTOMATED ORIGINATIONS FOR TIME, WEATHER, AND NEWS
4 THREE SIGNALS IMPORTED BY MICROWAVE
5 DENSITY OF 90 HOMES PER MILE OF CABLE
6 NO COPYRIGHT FEES FRANCHISE FEE = 5 PERCENT OF TOTAL REVENUES
7 SUBSCRIBER REVENUES = $67 92 PER SUBSCRIBER PER YEAR
8. UNDERGROUND CABLE = 5 PERCENT OF TOTAL PLANT
9 POLE AND TOWER RENTAL = $250/MILE

[c] ASSUMES IN ADDITION TO ALL OF ABOVE·

10 LIVE ORIGINATIONS AT AN ANNUAL OPERATING COST OF $47,300
11 ANNUAL CERTIFICATION BY FCC
12 FCC FEE OF $35 PLUS $0 30 PER SUBSCRIBER
13 PUBLIC SERVICE CHANNEL COSTS OF $4,875 PER YEAR

tration, approximately the industry averages for 1972.[19] Once more, we assume microwave importation of three signals and automated originations prior to the imposition of the 1972 FCC Rules, but we also estimate operating costs under C-M parameters for conditions after the imposition of the Rules.[20] These two sets of estimates appear next to C-M's own estimates for a skeletal system—no originations nor signal importation—for systems outside the top 100 markets prior to passage of the 1972 Rules.

All three series demonstrate considerable efficiencies of large size with operating costs per subscriber declining by from 16 to 28 percent as system size trebles from 5,000 to 15,000 subscribers at maturity. The operating results of the four MSO's cited above do not bear out the existence of such large scale economies—perhaps because multiple system owners can obviate indivisibilities in the management and promotion of systems through centralization of these functions. Whatever the explanation, we do not feel confident in using C-M operating cost data for small systems nor for very large ones. For systems of about 10,000 subscribers, however, their results appear plausible.

Other observers of the CATV industry have suggested that C-M cost estimates may be as much as one-third too high, but they provide no basis for such a judgment.[21] Our conclusion from evidence mustered from MSO's serving approximately one-sixth of all CATV subscribers in 1970 is that they have overestimated the economies of scale in system operation. Further corroboration of this conclusion may be found in another paper by one of the present authors[22] which estimates the projected growth in cash flow econometrically from a sample of system sales in the 1971–1972 period. The results of that study demonstrate no significant economies of large scale of cable system operation.

---

[19] *Weekly Television Digest,* September 25, 1972, and October 2, 1972; data for 50 largest multiple system owners, serving 70 5 percent of nation's 6.5 million subscribers.

[20] The assumptions employed in the latter derive from Mitchell and Smiley [6], Appendix.

[21] See Noll, Peck, and McGowan [8].

Finally, common sense leads us to reject the C-M results for smaller systems. In 1972, 83 percent of the nation's cable systems had fewer than 3,500 subscribers.[23] If the estimates in the last column of Table 6 were accurate, we would expect all systems of 3,500 or fewer subscribers to be courting bankruptcy as a result of the 1972 FCC Rules, for even a system as large as 3,500 subscribers would be realizing only $5.71 per year in cash flow from each $67.92 collected from subscriber fees. Given average investment of $200 per subscriber[24] this cash flow would hardly be sufficient to cover fixed charges and investors' demands for return on equity. But no such bankruptcy wave exists, undoubtedly because cash flow is a much larger proportion of revenues for these smaller systems than C-M would predict.

☐ **Capital costs.** Since cable television is an extremely capital-intensive activity, assumptions about the magnitude and timing of capital expenditures are crucial to any rate of return calculations. There are three major reasons why both C-M and Mitchell's capital expenditures lead to a downward bias in calculated profitability.

(1) They assume that underground cable percentages are far above actual and prospective underground percentages in the most dense markets.
(2) They assume that the entire plant is built at the beginning of the first year.
(3) They fail to account for a less expensive method of construction in dense middle markets.

We shall take issue explicitly with the first two of these assumptions, citing data collected from middle-market systems either in operation or under construction at the present time and a recent Rand study of the prospects for cable in the Dayton-Miami Valley area and the projected temporal pattern of construction for this system.[25]

Underground construction. C-M's and Mitchell's assumptions about the extent of underground construction in the largest markets are unrealistic. They predict that the average percent of underground construction will vary by market and by proximity to the middle of the market. A tabulation of these percentages appears as Table 7.

---

[22] Crandall [4].

[23] See [11], 1972–1973 edition.

[24] See Andersson [1].

[25] See Johnson et al. [5].

TABLE 7

COMANOR-MITCHELL AND MITCHELL'S ASSUMPTIONS ON UNDERGROUND CONSTRUCTION PERCENTAGES

LOCATION	MITCHELL		COMANOR-MITCHELL	
MARKET	MIDDLE SYSTEMS	EDGE SYSTEMS	MIDDLE SYSTEMS	EDGE SYSTEMS
TOP 50	20%	10%	20%	5%
51-100	10%	5%		
101+	5%	5%	5%	—

A more realistic assessment of the probable extent of underground construction in the center of the largest markets—where underground construction is likely to be most important—may be gleaned from the data in Table 8 which are derived from direct telephone contacts with managers of cable systems operating in the center of the largest 50 markets, excluding New York City.

TABLE 8

UNDERGROUND CABLE PERCENTAGES -- CORE CITIES IN TOP 50 MARKETS

CITY	PERCENT UNDERGROUND CURRENTLY	PROSPECTIVE PERCENTAGE UNDERGROUND
AKRON, OHIO	LESS THAN 5%	LESS THAN 5%
COLUMBUS, OHIO	LESS THAN 5%	LESS THAN 5%
BUFFALO, N Y	LESS THAN 1%	PERHAPS AS MUCH AS 6%
ST PETERSBURG, FLA	0%	0%
ATLANTA, GA	0% (EXCEPT LEADS TO APARTMENT COMPLEXES)	0%
SEATTLE, WASH	10-15%	—
ALBANY, N Y	—	APPROXIMATELY 2%
CHARLOTTE, N C	10%	10-11 5%
ANDERSON, S C	LESS THAN 1%	LESS THAN 1%
ASHEVILLE, N C	0 1%	LESS THAN 0 5%
SAN FRANCISCO, CALIF.	16%	LESS THAN 12%
OAKLAND, CALIF	23%	—
LOS ANGELES, CALIF.	23%	—
SAN JOSE, CALIF	23%	—
WINSTON-SALEM, N C	5%	UP TO 7-8%
SALT LAKE CITY, UTAH	0%	0%

SOURCE TELEPHONE CONVERSATIONS WITH SYSTEM MANAGERS AND ENGINEERS, 1973

Outside California and New York City, underground cable percentages rarely exceed 5, and in California—where new housing construction is especially important—underground cable costs are often much lower than those specified by both earlier studies because the television cable is buried at the same time that other utility lines are laid. Thus, there is little sound data from actual or prospective cable systems that underground percentages will "typically" average 10 or 20 in even the largest markets. In cities such as New York, Washington, or San Francisco, where topography and other unusual circumstances dictate burying cable at great expense, higher subscriber fees will be paid by television homes. But these cities should not be used as the model for "typical" systems in calculating rates of return. To do so would bias the results severely.

Timing of capital expenditures. A second assumption which leads both earlier studies to underestimate the internal rate of return on cable systems is the assumption that the entire plant is constructed at the beginning of the first year. In virtually every large system, construction is phased out over more than one year, and in many completion requires 3 or more years. In their study of the Dayton-Miami Valley system, the Rand staff assumed that the distribution system would be built in a 3-year period, with 21 percent completed in the first year, 44 percent built in the second year, and 35 percent in the third year.[26]

Alternative construction methods. C-M and Mitchell assume that large underground cable percentages will be achieved at a cost of $15,480 per mile because they implicitly assume that large cities will allow and even require cable companies to trench and lay conduit throughout the core city area. In fact, this is not likely to occur in many large cities because of the costs and discomforts created by the trenching and filling process. Many cities require instead that the telephone or electric utility company provide conduits for other purposes. These conduits are laid when other utility lines are laid, and cable operators may be required to use them. In Boston, one of Mitchell's examples for extensive underground requirements, city engineers are emphatic in asserting that new trenches will not be dug by a cable television company, requiring instead that cable be placed in existing utility ducts.

In other cities, such as Seattle, San Jose, and Los Angeles, underground cable is laid at the same time utility companies bury their cables. This leads to much lower costs than independent *de novo* construction hypothesized by Mitchell. Even where underground construction is undertaken independently, the cost of laying cable can vary enormously. Oakland is laying 38 channels of cable underground at an estimated cost of $19,000 per mile while San Francisco, across the bay, is encountering costs of up to $50,000 per mile and more. These differences are reflected in different monthly charges—$4.45 for Oakland and $6.25 for San Francisco.

Where cable is simply strung through underground ducts, the cost per mile may even be lower than aerial construction. One study conducted by the Stanford Research Institute[27] found that this type of construction cost only $3,000 per mile, or at least 25 percent less than aerial construction.[28] Maintenance costs may be greater for this alternative, however, since major repairs or alterations may require the assistance of other utility companies who share the same ducts.

In our calculations in the next section, we shall make the reasonable assumption that the underground percentage in the typical system is 5 percent. Moreover, we shall phase the initial (and subsequent replication of) investment over 2 years, a pattern which the Rand Dayton study uses in each of the sections of its enormous prospective system. Unfortunately, we cannot present very firm data on the percentage of cable which will be laid in utility ducts. Therefore, we utilize C-M's data on aerial and underground costs per mile.

---

[26] *Ibid.*

[27] See Stanford Research Institute [10].

[28] More recent estimates are considerably higher, running as high as $8,000 per mile, but are still considerably less than new trenching.

4. The revised estimates of profitability

■ In this section, we provide our estimates of the prospective profitability of various cable systems under the new FCC Rules. We utilize the following assumptions, many of which have been discussed above.

☐ **Subscriber penetration.** (Subscribers/Homes Passed at Maturity):

(1) For all systems outside the top 100 markets—0.653.
(2) For systems in the top 50 markets, alternatively 0.648, 0.50, and 0.40.
(3) For systems in markets 51 to 100, alternatively 0.653, 0.50, and 0.40.

☐ **Maturation.** We follow Mitchell's path requiring that 30 percent of subscribers be enrolled by the end of year 1, 55 percent by the end of year 2, 80 percent by the end of year 3, 95 percent by the end of year 4, and 100 percent by the end of year 5.

☐ **Subscribers.** The number of subscribers increases at an annual rate of 2 percent in years 6 through 60 as a result of new household formation.[29]

☐ **Costs and revenues.** All operating costs and revenues are discounted from the middle of each year. Capital costs are discounted from the beginning of the year in which they are incurred.

☐ **Investment.** All distribution plant investment is spread over 2 years—50 percent in the first year, 50 percent in the second year.

☐ **Underground.** Underground cable percentage is 5 percent in all markets.

---

[29] Mitchell has suggested to us that this assumption may be unjustified. He utilizes a 2-percent growth rate to reflect household formation or simply new subscribers during years 6 through 10, but thereafter he keeps subscriber levels at their tenth-year level. It is possible that a 2-percent growth rate throughout our sixty-year period is too liberal and that perhaps 1 percent would be a more accurate projection of the next sixty years in most markets. But this additional 1-percent growth per annum in subscriber levels has only a minor effect upon estimated profitability.

☐ **Capacity.** Twenty-channel capacity for all systems.

☐ **Exclusivity effects.** Two additional independents imported via three microwave hops each as standby capacity to substitute for blacked-out independents.

☐ **Origination.** All systems have standard origination as defined by C-M.

☐ **Annual revenue per subscriber.** $67.92 which grows at a rate of zero percent, 2 percent, and 4 percent in three different calculations.

☐ The important differences between this set of assumptions and those utilized by Mitchell are concentrated in the first and last items. We allow subscriber penetration at maturity to assume three values—(1) 0.653, depending upon the market category, precisely the values predicted for 1987 by multiple system operators canvassed by the N.C.T.A.;[30] (2) 0.50—the current industry average for all markets;[31] and (3) 0.40—a reflection of Park's estimate for the typical large market system which performs significant originations.[32] In addition, we demonstrate the effect of allowing revenues per subscriber to grow at rates of 0 percent, 2 percent, or 4 percent per annum, reflecting different assumptions concerning the importance of future ancillary services such as leased channels.

Since we use the C-M operating cost assumptions, we base all of our projections upon a 10,000 subscriber system, a scale of operation for which the C-M assumptions generate reasonable cost projections. All calculations are based upon an assumed 60-year life. Finally, we utilize the proposed copyright fee schedule which appears in S.1361, known as the (Sen.) "McClellan" proposal. This fee schedule is reproduced as Table 9 below.

TABLE 9

"McCLELLAN" COPYRIGHT FEE SCHEDULE (FEE AS A % OF GROSS REVENUE)

PERCENTAGE OF FIRST $160,000 IN REVENUES	1 0%
PERCENTAGE OF REVENUES IN EXCESS OF $160,000 UP TO $320,000	2 0%
PERCENTAGE OF REVENUES IN EXCESS OF $320,000 UP TO $480,000	3 0%
PERCENTAGE OF REVENUES IN EXCESS OF $480,000 UP TO $640,000	4 0%
PERCENTAGE OF REVENUES IN EXCESS OF $640,000	5 0%

The remaining assumptions are rather self explanatory, differing from Mitchell's and C-M's to only a minor degree due to considerations delineated above.[33] The sensitivity of varying the important assumptions may be gleaned from Table 10, which

---

[30] Andersson [1]. The year 1987 is chosen because we are attempting to make long-run projections for a nascent industry. According to this N.C.T.A forecast, only slightly more than half of the homes in the top 100 markets will have access to cable by 1987.

[31] *Weekly Television Digest,* September 25, 1972, and October 2, 1972.

[32] Park [2] uses this estimate and a similar justification for forecasting 0.40 penetration for the Dayton, Ohio, area.

[33] Among these are the spreading of capital investment over two years, the use of standby signals to combat the effect of the FCC exclusivity rules, a maximum of 5 percent of cable laid underground, $67.92 in fees per subscriber in the first year (rather than $64.60 or $69.00 utilized by Mitchell and C-M); and minor differences in employee benefits, office rentals, and house drops because of difficulties encountered in reproducing the C-M results from their assumptions stated in Appendix 3.

TABLE 10

SENSITIVITY OF RATE OF RETURN CALCULATION TO CHANGES IN PARAMETERS (NO COPYRIGHT FEES)

	25,000 SUBSCRIBER SYSTEM—MIDDLE OF MARKETS 1 TO 50				
	CASH FLOW DISCOUNTED AT BEGINNING OF YEAR	CASH FLOW DISCOUNTED IN MIDDLE OF YEAR INVESTMENT COMPLETED IN YEAR 1		CASH FLOW DISCOUNTED IN MIDDLE OF YEAR INVESTMENT COMPLETED IN 2 YEARS	
	UNDER - GROUND =20%	UNDER - GROUND =20%	UNDER - GROUND =5%	UNDER - GROUND =20%	UNDER - GROUND =5%
1 MITCHELL'S CALCULATION REVENUE=\$64 60 PENETRATION AT MATURITY=0 272	10 4%	11 7%	13 6%	12.1%	14 2%
2 OUR CALCULATION REVENUE=\$64 60 (ADDITIONAL MICROWAVE HOPS) PENETRATION AT MATURITY=0 278		11 8%	13 8%	12 1%	14 1%
3 OUR CALCULATION REVENUE=\$67 92 PENETRATION AT MATURITY=0 278		13 0%	15 1%	13 4%	15 6%
4 OUR CALCULATION REVENUE=\$67 92 PENETRATION AT MATURITY=0 648					28 5%
5 OUR CALCULATION REVENUE=\$67 92 $(1.02)^{t-1}$ PENETRATION AT MATURITY= 0 648					32 8%
6 OUR CALCULATION. REVENUE = \$67 92 $(1\ 04)^{t-1}$ PENETRATION AT MATURITY=0 648					36 8%

provides a sample calculation for a 25,000-subscriber system in the center of the top 50 markets. Mitchell projects a 10.4 percent rate of return for such a system while our algorithm generates an estimate of 11.7 percent for the same conditions. The difference is due in large part to our assumption of a 2-percent annual growth rate in subscribers from year 6 through year 40, reflecting new household formation. Mitchell allows subscriber levels to grow at 2 percent per year only in years 6 through 10. Otherwise, Table 10 provides a rather clear indication that more optimistic forecasts of subscriber levels and of revenue growth have an enormous effect upon projected rates of return.

The range of rates of return which emerge from our various assumptions is presented in Tables 11 through 15. Each table contains the range of predictions for the market categories delineated by Mitchell. Rates of return are calculated with and without copyright fees and for the entire range of assumed penetration and revenue growth.

In the top 50 markets, our representative system earns more than the cost of capital at a mature subscriber penetration of 64.8 percent of homes passed even with no revenue growth. More pessimistic penetration assumptions provide rates of return which are below the cost of capital if we continue to assume zero revenue growth, but once we allow gross revenues to grow at a modest 2 percent per year, the rates of return rise rather strongly. With the

TABLE 11

RATES OF RETURN IN 10,000 SUBSCRIBER SYSTEMS IN THE MIDDLE OF THE TOP 50 MARKETS

		NO COPYRIGHT FEES	COPYRIGHT FEES		
		SUBSCRIBER PENETRATION AT MATURITY= 0.648	SUBSCRIBER PENETRATION AT MATURITY=		
			0 648	0 50	0 40
REVENUE PER SUBSCRIBER	=$67 92	18 61	17 65	15 32	13 28
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	23 47	22 54	20 34	18 43
REVENUE PER SUBSCRIBER	=$67.92 PLUS 4% ANNUAL GROWTH	27 63	26 68	24 47	22 55

OTHER VARIABLES
DENSITY 200 HOMES/MILE
IMPORTED SIGNALS 2 INDEPENDENTS, 1 DUPLICATE NETWORK, 1 EDUCATIONAL
MICROWAVE HOPS 15
UNDERGROUND CABLE 5%
INVESTMENT PERIOD FOR DISTRIBUTION SYSTEM. 2 YEARS

TABLE 12

RATES OF RETURN IN 10,000 SUBSCRIBER SYSTEMS IN THE EDGE OF THE TOP 50 MARKETS

		NO COPYRIGHT FEES	COPYRIGHT FEES		
		SUBSCRIBER PENETRATION AT MATURITY= 0.648	SUBSCRIBER PENETRATION AT MATURITY=		
			0 648	0 50	0 40
REVENUE PER SUBSCRIBER	=$67.92	16 30	15 40	12 95	9 86
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	21 22	20.37	18 08	15 41
REVENUE PER SUBSCRIBER	=$67 92 PLUS 4% ANNUAL GROWTH	25 33	24 47	22 17	19 55

OTHER VARIABLES·
DENSITY 150 HOMES/MILE
IMPORTED SIGNALS 2 INDEPENDENT, 1 DUPLICATE NETWORK, 1 EDUCATIONAL
MICROWAVE HOPS 15
UNDERGROUND CABLE. 5%
INVESTMENT PERIOD FOR DISTRIBUTION SYSTEM 2 YEARS

McClellan copyright fee schedule in force, all 10,000 subscriber systems in the top 50 markets are able to cover their cost of capital with only 50 percent penetration and 2-percent revenue growth. The assumption of 4 percent revenue growth allows these systems to more than cover their cost of capital of 15 to 16 percent even with only 40 percent of homes subscribing at maturity and McClellan copyright fees being paid.

Similar conclusions apply for markets 51 to 100. Only if we assume that subscriber penetration at maturity is as low as 40

TABLE 13

RATES OF RETURN IN 10,000 SUBSCRIBER SYSTEMS IN THE MIDDLE OF MARKETS 51 to 100

		NO COPYRIGHT FEES	COPYRIGHT FEES		
		SUBSCRIBER PENETRATION AT MATURITY=	SUBSCRIBER PENETRATION AT MATURITY=		
		0 653	0 653	0 50	0 40
REVENUE PER SUBSCRIBER	=$67 92	16.39	15 48	12 95	9 86
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	21 30	20 45	18 08	15 41
REVENUE PER SUBSCRIBER	=$67 92 PLUS 4% ANNUAL GROWTH	25 42	24 55	22 17	19 55

OTHER VARIABLES·
DENSITY 150 HOMES/MILE
IMPORTED SIGNALS 2 INDEPENDENT, 1 EDUCATIONAL
MICROWAVE HOPS 15
UNDERGROUND CABLES 5%
INVESTMENT PERIOD FOR DISTRIBUTION SYSTEMS 2 YEARS

TABLE 14

RATES OF RETURN IN 10,000 SUBSCRIBER SYSTEMS IN THE EDGE OF MARKETS 51 TO 100

		NO COPYRIGHT FEES	COPYRIGHT FEES		
		SUBSCRIBER PENETRATION AT MATURITY=	SUBSCRIBER PENETRATION AT MATURITY=		
		0 653	0 653	0 50	0 40
REVENUE PER SUBSCRIBER	=$67.92	14 82	13 96	10 42	8 40
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	19 79	18 99	15 88	14 09
REVENUE PER SUBSCRIBER	=$67.92 PLUS 4% ANNUAL GROWTH	23 88	23 07	20 02	18 20

OTHER VARIABLES
DENSITY 125 HOMES/MILE
IMPORTED SIGNALS 2 INDEPENDENTS, 1 EDUCATIONAL
MICROWAVE HOPS 15
UNDERGROUND CABLE 5%
INVESTMENT PERIOD FOR DISTRIBUTION SYSTEMS 2 YEARS

percent of homes passed, and that revenues per subscriber do not grow at all, do we obtain rates of return below 15 percent. Mildly optimistic forecasts of 50-percent penetration and 2-percent revenue growth allow McClellan copyright fees to be paid without endangering the financial condition of the typical cable system.

The smallest markets provide sufficient net revenues to cover copyright fees for a revenue growth rate of 2 percent. At higher growth rates, there are considerable profits in excess earned by all systems.

TABLE 15

RATES OF RETURN IN 10,000 SUBSCRIBER SYSTEMS IN MARKETS 101+

MIDDLE MARKET SYSTEMS		SUBSCRIBER PENETRATION AT MATURITY=0.653	
		NO COPYRIGHT FEES	COPYRIGHT FEES
REVENUE PER SUBSCRIBER	=$67.92	16 66	15 77
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	21 49	20 64
REVENUE PER SUBSCRIBER	=$67 92 PLUS 4% ANNUAL GROWTH	25 56	24 70
EDGE-MARKET SYSTEMS			
REVENUE PER SUBSCRIBER	=$67 92	13 52	12 69
REVENUE PER SUBSCRIBER	=$67 92 PLUS 2% ANNUAL GROWTH	18 62	17 86
REVENUE PER SUBSCRIBER	=$67.92 PLUS 4% ANNUAL GROWTH	22 72	21 96

OTHER VARIABLES

DENSITY = 125 HOMES/MILE FOR MIDDLE-MARKET SYSTEMS, 100 HOMES/MILE FOR EDGE-MARKET SYSTEMS
IMPORTED SIGNALS· 1 INDEPENDENT, 1 EDUCATIONAL
MICROWAVE HOPS 12
UNDERGROUND CABLE 5%
INVESTMENT PERIOD FOR DISTRIBUTION SYSTEM· 2 YEARS

5. Conclusions

■ Based on the foregoing calculations, our conclusions on the ability of cable systems to develop in major markets and to pay copyright fees is much more sanguine than that offered by C-M and by Mitchell. The difference derives largely from the dubious assumptions made by both in their analyses—assumptions wholly inconsistent with N.C.T.A. members' own predictions and recent behavior in cable system sales. Once the low penetration rates and conservative revenue estimates are replaced by more reasonable industry estimates, rates of return rise strikingly, even in the face of copyright fee payments.

While we do not have sufficient evidence to provide an accurate prediction of the long-run profitability of cable, we do demonstrate that current estimates of prospective profitability are quite sensitive to variations in the assumptions employed, especially those concerning future revenue sources and prospective subscriber penetration. We have not identified prospective sources of cost reductions, and in this respect our profitability estimates may be conservative. For example, we have given no consideration to prospective technological or organizational improvements such as signal distribution via microwave common carriers or by satellite.

Implicit in the above conclusions is the belief (shared by many system owners) that new sources of revenues from ancillary services such as "premium television" will be substantial and will have a salutary effect upon cable demand. Unfortunately, however, the largest variable in prospective cable profitability, notwithstanding recent accomplishments, continues to be the development of future regulatory policy by the FCC and the Congress. To the extent regulators act to restrict the development of new revenue sources, our assumptions of revenue growth would not be justified.

Finally, any exercise such as that attempted in this paper is in no small part an attempt to estimate the degree of slack in local cable regulation. If subscriber fees are continuously forced downward by municipal authorities in response to new sources of ancillary revenues, rates of return may simply approximate the cost of capital to cable owners. Alternatively, if municipalities view the potential of profits in excess of necessary capital charges principally as a source of supporting local educational, informational, or political services, profitability estimates will have to be reduced considerably. It is apparent that investors have not expected regulators to behave in such a fashion, and our estimates reflect rather similar expectations.

References

1. ANDERSSON, D. *The CATV Industry—Current Subscriber Penetration, Projected Subscriber Growth, Capitalization, by 5-Year Periods, 1972–1992.* National Cable Television Association, 1972.
2. BAER, W. AND PARK, R. E. Paper #2 in L. L. Johnson, *et al.*, eds., *Cable Communications in the Dayton Miami Valley. Basic Report,* The Rand Corporation Report No. R943KF/FF, January 1972 [5].
3. COMANOR, W. AND MITCHELL, B. "Cable Television and the Impact of Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 154–212.
4. CRANDALL, R. "Estimation of the Profitability of Cable Television from Recent System Acquisition Prices." Manuscript, 1973.
5. JOHNSON, L. L. ET AL., EDS. *Cable Communications in the Dayton Miami Valley: Basic Report.* The Rand Corporation Report No. R943KF/FF, January 1972.
6 MITCHELL, B. IN ASSOCIATION WITH SMILEY, R. "Cable, Cities, and Copyrights." *The Bell Journal of Economics and Management Science,* Vol. 5, No. 1 (Spring 1974), pp. 235–263.
7. MYERS, S. C. "The Application of Finance Theory to Public Utility Regulation." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 58–97.
8. NOLL, R., PECK, M., AND MCGOWAN, J *Technical Memorandum to the Sloan Commission on Cable Television,* 1971.
9. PARK, R. E. "Prospects for Cable in the 100 Largest Television Markets." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 130–150.
10. STANFORD RESEARCH INSTITUTE. *Business Opportunities in Cable Television.* March 1970.
11. *Television Factbook,* 1970–1971, 1972–1973.
12. U. S. FEDERAL COMMUNICATIONS COMMISSION. *Final Cable Television Decision,* Docket 18397, February 2, 1972.
13. WEINBERG, G. *Cost Analysis of CATV Components.* Report prepared for the Office of Telecommunications Policy, Executive Office of the President. Bethesda, Md.· RMC, Incorporated, June 1972.

Cowling's *Market Structure and Corporate Behavior*

Reviewed by

Leonard W. Weiss

Professor of Economics
University of Wisconsin—Madison

■ This is a collection of seven essays on assorted topics in industrial organization.[1] Most are reports on research in process. The theoretical analysis is often impressive. Most of the papers end with empirical studies that are preliminary and, in a number of cases, unconvincing.

In the first piece, Hindley reviews the literature on the causes and effects of mergers. He emphasizes the work of Gort and Dennis Mueller and merely bows to traditional notions about mergers for monopoly or economies of scale and ignores short-run speculative motives. He emphasizes mergers to transfer capital or specialized skills or to increase growth or head off takeovers where managements are growth-maximizers. He notes previous evidence that merging firms are less profitable than either their component predecessors or nonmerging firms and finds it unconvincing because merger may be an attempt to remedy an even poorer prospective performance. He cites his own evidence that a large number of poor performance, widely held firms survive in spite of widespread mergers. This, he feels, is conclusive evidence that the threat of takeover is insufficient to make managers into present value maximizers. It is also inconsistent with growth maximization since a growth maximizer would put a higher value on any firm, efficient or not, than capital markets do.

Kuehn tries to evaluate the motives of those firms engaged in takeovers in Britain during 1957 to 1970. He works out expected differences between takeover raiders and other firms if the raiders are, alternatively, growth or profit maximizers and if the other firms are, alternatively, maximizers of profits or growth, or of the easy life (safety) or "sleepy firms." The variables considered are growth rates, profit rates, stock market valuation ratios, and income retention ratios. The raiders grew faster and had higher valuation ratios and lower retention ratios than other firms listed on the London Stock Exchange. These results best fit the *a priori* expectations for growth-maximizing raiders and safety minded nonraiders. This is not a well-tested hypothesis, however, since any statistically significant differences would have supported one of his hypotheses, and his tests are entirely univariate. A striking fact reported is that 47 percent of all publicly listed British commercial and industrial corporations were taken over between 1957 and 1970.

---

[1] Keith Cowling, ed., *Market Structure and Corporate Behavior* (London: Grey Mills Publishing Company, 1972).

Samuels and Chester study births, deaths, and growth rates in a sample of listed British firms in the period 1960 to 1969. As usual, there is little correlation between size and growth rates, and births and deaths are concentrated in the small size classes. They found deaths far more frequent than births, but this may be due to a bias in their sample. They used a stratified random sample for births and survivors, but counted all deaths observed among cases drawn and rejected in finding the other two groups. They found increasing concentration using the variance of logs of firm size among surviving firms. This is inevitable if Gibrat's law holds. It is of doubtful value because the variance of logs is so heavily dependent on large numbers of small firms. Moreover, takeovers by foreign firms are counted as deaths though the acquired firms often remained in the market. Their results to date are neither startling nor very convincing.

Jacquemin and Thisse review the applications of optimal control theory to such issues as pricing where buyers respond with a lag, the effects of entry barriers on the time paths of price and market share, advertising strategies where ads have long-lasting effects, and research and development strategies. The usual prediction is that present value maximizers will forego short-run profits in early periods to attain subsequent profits. The entry barriers and advertising analyses may well be sufficiently developed to permit empirical tests. The price and R and D analyses seem far short of that point. The R and D analysis, by ignoring diseconomies of scale to research and by assuming significant externalities for both firm and industry, arrives at the doubtful conclusion that optimal levels of R and D will occur only in a pure monopoly with R and D subsidies.

Cowling evaluates advertising intensity decisions of British firms by using the Dorfman-Steiner model. He points out that if buyers respond to prices with a lag, then the Arrow-Nerlove analysis may reduce to the Dorfman-Steiner formulation. His empirical tests consist of regressions based on pooled cross sections in the British auto, tractor, margarine, coffee, and toothpaste industries for the years 1957 to 1968. His study yields brand advertising elasticities (mean for five products of 0.351) and brand price elasticities (mean of —2.605), numbers that are rarely seen. His results for nondurables approximate the Dorfman-Steiner rule, but ads appear to be underutilized in autos and tractors. He suggests as reasons that he may overstate advertising elasticities for durables because ads shift demand among model years as well as among brands or because heavy advertising for durables coincides with large design changes for which he did not control. Alternatively, the result may be due to recognized interdependence in the auto and tractor industries.

Cable attempts to relate advertising intensity to market structure. His theory is much more sophisticated than those of previous studies in this area. He starts with the Dorfman-Steiner and Arrow-Nerlove formulations and, given Cournot assumptions, predicts an inverted *U*-shaped advertising function with respect to the

Herfindahl index. The alternative assumption of recognized interdependence yields no relationship between concentration and advertising but an inverted-*U* relationship if they collude on price only. He expects advertising elasticity to increase with the number and turnover of buyers and to decrease with consumer search and the number of brands. Finally, he expects demand growth to increase advertising elasticity and decrease the danger of retaliation. He works with products chosen to limit the effects of these variables: 26 consumer nondurables sold through British food and drug stores to many households on a continuing basis. His dependent variables were, alternatively, the advertising sales ratio and "goodwill," the depreciated stock of advertising investments using a 30 percent per year depreciation rate. Goodwill yielded the best fit. The coefficient for the Herfindahl index was always significantly positive and that for the index squared significantly negative. The predicted maximum goodwill occurs at $H = 0.4$ (i.e. where the number of equal-sized firms is about 2.5). The three-firm concentration ratio also had a significant positive effect but its quadratic form did not fit well. The various product characteristics had equivocal effects. A dummy for products closely associated with such basic values as sex, children, health, beauty, and status had a strong positive effect as did the number of brands, but total sales, sales growth, and brand turnover had insignificant effects. Though he describes more ambitious work in progress, that already done is more sophisticated and convincing than any previous work on the subject.

Williamson and Bhargava's paper is a useful brief summary of their recent and continuing work on the internal organization of the firm. They distinguish "unitary" firms, where all decisions are made by central staff, "holding companies," where divisions are largely autonomous, and "*M*-form" corporations, where divisions have autonomy on short-run, operating decisions but where central staff makes strategic decisions. A unitary structure is prescribed for small firms. Holding companies are little better than mutual funds and may be worse because of the inflexibility of their portfolios and their tendency to bail out failures, thus creating slack. But they feel the *M*-form corporation is superior to independent firms connected through capital markets, because its superior information on divisional performance permits it to design rewards to induce profit maximization, to allocate capital on the basis of marginal instead of average profit prospects, and to judge risks more realistically. It can structure interdivisional transactions and competition to allow for intrafirm externalities. The *M*-form structure fails when central staff become regularly concerned with divisional operating decisions. They view the *M*-form company as a major innovation of the interwar period which was first widely applied in the late 1940s and became common in the 1950s and 1960s. They expect increased competition and closer approaches to profit maximization as *M*-form firms become more common. Many points made are familiar, but this seems to be their most succinct statement.

Yarrow studies the determinants of executive compensation. His theory is more sophisticated than those of his predecessors. He uses a neoclassical model where the marginal product of managers depends on such complementary inputs as capital stock, labor force, and staff, and where profits rise with management quality. He also cites an organization theory model where chief executive salary depends on the number of steps in an hierarchical organization. Using samples of 72 to 78 U. S. firms for 1963, 1965, and 1968, he regresses chief executives' salaries on various indexes of size (revenue, employees, assets, or equity) plus general, administrative and sales staff and the lagged rate of return on equity. Size and staff have significant positive effects, with assets or equity size yielding the best fits. Profit rates have insignificant positive effects in general cross sections but significant effects in producer goods industries when industry dummies are introduced. He feels that the effect of profit rates on salary varies with the industry because of the differing effects of other inputs on management productivity with differing production processes. His study suffers from insufficient observations, a lack that could be remedied by a visit to the SEC library. His results support both the neoclassical and hierarchical organization theories, although the latter is incomplete.

Phillips tests his familiar thesis that conduct variables belong in a profit determination model. He uses a remarkable set of data on the presence and effectiveness of price-fixing agreements from a survey of British industrial buyers in 1953–1956. Price fixing was reported for 243 out of 1200 British trade associations. It occurred in 26 of 71 industries studies. He regresses average price-cost margins on an arbitrary numerical index based on coverage of the price-fixing agreements and the survey estimates of the effectiveness of agreements. Other variables were concentration, average plant size, advertising intensity, industry growth, a concentration-advertising interaction, and a producer good dummy. The collusion variables had expected signs, but were insignificant. Concentration and advertising had significant positive effects, but in a quadratic form margin reached a maximum at a three-firm concentration ratio of 48 to 76 depending on the regression used, and the positive effect of advertising disappeared where concentration exceeded 60 to 71, depending on the regression. Plant size and growth had unexpected negative effects on margins. Price agreements were more common but less effective, the larger the number of trade associations in an industry. He interpreted the number of trade associations as an index of heterogeneity. Phillips' data is probably as good as any we shall ever get on formal collusion. The conventional wisdom that conduct is of secondary importance seems to be supported.

Overall, this book is uneven. It is a stimulating source of researchable hypotheses, partly because of the high level of analytical sophistication in many studies, but the actual empirical work involved is often incomplete.

Posner's *Economic Analysis of Law*

Reviewed by

Peter A. Diamond

Professor of Economics
Massachusetts Institute of Technology

■ Richard Posner has written an exceptionally interesting book.[1] It is a wide ranging discussion of the impact of the legal system on the allocation of resources by the market and of economic aspects of the behavior of participants in the legal system, written as a text for law students knowing little or no economics by a nonmathematical Chicagoan. (Of course, it is only the absence of mathematics that I hold against the author.) I shall briefly lay out the range of analysis and variety of audiences who will find this book interesting, some limitations of the absence of mathematics, and some examples of the tone of the book. Then, in a more leisurely fashion, I shall discuss contract law to show an example of the type of analysis in the book.

This is a textbook for law students, intending to teach economic principles while conducting a survey of the rules and institutions of the legal system, summarizing, and extending the applications of economics to law. I expect that it does its job well and would be an interesting book to teach or study from in this sort of course, for it is very well written and interesting in content. As an economist, I naturally believe that the book would be better for law students if it taught a little more economics, particularly noting which assumptions play major roles in the analysis. (I also suspect that some discussion of the different nature of marginal and lumpy decisions would be valuable, i.e. the complications arising from nonconvexities.) The text would also serve very well for a law and economics course for economics students (either advanced undergraduates or graduate students), although the more conventional economics subjects, such as taxation and regulation, would need considerable expansion, or better, omission since they are already a standard part of economics curricula. While this is basically a text, it is also a monograph in a field of growing interest to economists, and I will review it primarily on that basis.

The scope of this book is exceptionally wide. It encompasses the traditionally analyzed areas of government intervention in the market (antitrust, regulation, taxation), the areas of great recent interest of the impact of the legal system on private use of the market (property rights, contracts, torts, crime control, antidiscrimination regulations) and of the economics of the functioning of the legal system (due process, civil and criminal procedure, law enforcement). With the hypothesis that the common law is seeking economic efficiency, it also seeks to explain judicial decision-making.[2] The style is lucid but unfortunately, to my taste, nonmathematical except for numerical examples and a few diagrams. Mathematics is both a method of exposition that makes some assumptions (or their importance) more apparent and a method of thought that makes some types of errors more difficult. For example, in the discussion of expenditures on litigation,[3] under the assumption that each party takes the litigation expenditures of the other as given, it is stated that increased optimism results in higher expenditures. But the perceived probability of success is a function of expenditures. For an upward shift in the function to increase expenditures, the derivatives must change appropriately since the first-order condition involves the derivative, not the level of probability.[4]

---

[1] Richard A. Posner, *Economic Analysis of Law* (New York: Little, Brown and Co., 1973).

Another example where mathematics might clarify comes in the discussion of free entry and profitability. With free entry and identical cost functions for all actual and potential firms, all firms have zero profits (i.e. just earn the normal return) in competitive equilibrium. When different firms have different cost functions, the marginal firm (with a continuum of firms) will have zero profits, but other firms may be earning positive profits (or rents) without encouraging further entry, since potential firms (which have not yet entered) have higher costs and cannot duplicate the profitable position of some of the existing firms. *A priori* I do not see any reason to assume that all potential entrepreneurs would start firms with identical cost functions. Particularly in a book where transaction costs play a major role, we would expect existing firms to have at least an informational advantage, vis-à-vis customers, relative to potential firms. Empirically, in particular industries, the supply of potential entrepreneurs may be such that the identical firm assumption is a good one. I do not believe it is a good assumption in all industries. This distinction is not drawn despite repeated use of a zero profit assumption (and the implied eventual bankruptcy of any firm that is not successfully profit maximizing). At least in organized crime one would expect an existing firm to have an advantage over potential firms, particularly if they are going to shoot it out for a (temporary?) monopoly position. (Do potential firms use potential bullets?)

---

[2] Of course, this hypothesis is not universally accepted. See, e.g., George Fletcher, "Fairness and Utility in Tort Theory," *Harvard Law Review* (January 1972).

[3] A dose of game theory might be helpful here, particularly in the discussion of the existence of equilibrium when each party is aware that litigation expenditures may induce a response.

[4] It seems to me that clearly distinguishing average from marginal ranks as an important economic lesson. Thus sloppiness on this point is a weakness in the book See, e.g. "he will stop spending [on an application for a broadcast license] at the point where, if he prevails, the difference between his expenditures and the value of the license will be large enough to compensate him for the risk," (p. 14). For the example in the text, expected profit of the plaintiff is $\Pi(E)D-E$, where E is expenditures, D damage award if the case is won, $\Pi(E)$ the probability of winning if E is spent. The first-order condition is $\Pi' D = 1$. E will increase with a shift in Π that increases Π' at the previous optimum (given suitable regularity conditions). Π could shift up and Π' shift down in the relevant range.

Now let us turn to a few quotations which will give the tone of the book more easily than description:

> There is abundant evidence that legislative regulation of the economy frequently, perhaps typically, brings about less efficient results than the market-common law system of resource allocation. [Footnote containing references omitted.][5]

> A newer view, which draws on economic analysis, argues that the structure of the administrative process is designed to increase political control over the process of legal regulation rather than to increase efficiency but that within the constraints imposed by the fundamentally political purpose of regulation the evidence is consistent with the hypothesis that the agencies, like most other organizations, are rational utility maximizers [6]

> The analysis in the preceding section assumes that judges make their decisions in accordance with the criterion of efficiency. But what is the linkage between the judges' self-interest and the promotion of efficient resource use?[7]

> Delay is not due to the fact that the demand for litigation is high and the amount of judge time limited. . . . An appropriately graduated system of surcharges for people desiring to have their cases heard promptly would [clear the market.][8]

> Without statutory protection against disinheritance, women could negotiate with their husbands for contractual protection. The statutory provision minimizes transaction costs.[9]

As this last quotation suggests, Posner does not examine the question of the effect of institutions on preferences and interpersonal relations. Rather, he uses the standard concept of Pareto optimality in terms of given preferences.[10] This mode of analysis is very well employed and pushed into new regions of uncertainty and transaction costs. One curious aspect of the structure of the book is the treatment of the transaction costs of the legal system. In much of the book, they are taken to be zero (particularly in the discussions of the correct measurement of damages to induce efficient behavior). Then, when the procedural system is directly discussed, the measure of damages which would be optimal for a frictionless legal system is assumed to hold. (This can, of course, be viewed as a version of how the legal system is currently trying to measure damages. Posner does discuss separately damages where the probability of apprehension is less than one.) Treating simultaneously the measurement of damages and the structure of procedures would be interesting.

---

[5] P. 329.

[6] P. 386.

[7] P 325.

[8] P. 355.

[9] P. 250.

[10] While the effect of legal rules on the incentive to settle disputes is examined, I was sorry to see no discussion of the very interesting analysis by Arthur Leff ("Injury, Ignorance and Spite—The Dynamics of Coercive Collection," *Yale Law Journal*, October 1970) of the dynamics of settlement and impact of legal rules on different types of litigants (wage earners-corporations, high volume litigants-infrequent litigants).

To see more of the flavor of this book and of this burgeoning field of research, let us consider one topic in more detail. I have selected the area of contracts.[11] Discussion of the role of contract law is greatly complicated by the presence of the contracting parties. By hypothesis they are (were, will be, or could easily be) negotiating. This prevents the simple analysis of tort law where the assumed high transactions costs prevent parties affected by externalities from negotiating to consider all of the social costs conventionally viewed by a general equilibrium model with externalities. Rather we must move into the fine structure of transactions to find any role for the law at all, other than dissuading outright cheating (and the costs induced by fear of cheating). There are three types of effects of the law which seem worth considering separately—adjusting allocation where the parties cannot or have not worked everything out, saving on their transaction costs in the process of working things out, and affecting information flows, and so, what they, are trying to work out. Put differently, the three types are where the parties are indeed strangers, where they are filling in details of an agreed bargain, and where they are determining the nature of the bargain.

In the first category Posner considers a contract where the delivery date stipulated omits the month. *A*, new to the industry, means next month; *B*, an old hand, knows and is following the custom of the industry that an omitted reference to the month means the current month. The court will interpret the contract as incorporating the customs of the industry. Posner points out that this encourages newcomers to master the language of the trade promptly, which presumably increases efficiency in the industry. One should also point out that it also affects the costs of entering a new field, and so, both long-run efficiency and income distribution. Imperfect communication can also take the form of faulty telegrams and garbled messages. In addition, where communication takes time, there are risks of changing conditions that must be borne in some division between the parties. By deciding who bears the risk (of increased costs, say, between transmission and receipt of an offer) the court can affect the efficiency of the bearing of this risk and the decisions affecting its magnitude. Since in these situations the parties are in imperfect communication, court imposed rules will generally determine where risks are borne (and so who adjusts to bear them).

An example in the second category considered by Posner is a quantity of wool with a latent defect that has not been examined by either the seller or the buyer. The buyer then uses the wool in manufacturing and so ruins the suits made from it. Efficiency calls for inspection by the party with lower inspection costs,[12] or no inspection if expected damages are lower than inspection costs. By trying to place liability on the cheaper inspector, the court can attempt to minimize the sum of expected inspection and damage costs. In the world of "transactors knowledgeable of the law," misplaced liability merely leads to a corrective transaction, shifting the liability. But corrective terms in the contract have transactions costs, so by getting liability right, the court permits a cheaper reliance on the common law to substitute for a more expensive explicit contract.[13] Partially restating the categorization of topics, Posner describes three economic functions, "maintenance of appropriate incentives" (i.e., putting risk where it is most cheaply borne), reduction of "the complexity and hence cost of transactions by supplying a set of normal terms"[14] and, not mentioned above, furnishing "prospective transacting parties with information concerning the many contingencies that may defeat an exchange, and hence to assist them in planning their exchange sensibly. The parties, through their lawyers, are guided around the pitfalls in the process of exchange revealed by the opinions in decided contract cases."[15]

---

[11] The selection reflects my own interests, of course, as well as the fact that Posner's analysis here is, to my knowledge, new and very interesting.

[12] These costs are not simply physical inspection costs but also reflect other transactions, for example whether self-protection against cheaters already called for some inspection by one of the parties. What is needed

Here we move into the third category, of making transactors more knowledgeable. The more difficult version of this is considering how people might have acted had they been more knowledgeable and considering how, if at all, legal rules applying both to carefully and carelessly worked out bargains should adjust for the careless. Let us look more closely at the role of information exchanges. Posner considers a contract induced by "an outright and calculated lie." In this case, he argues "the presumption that a contract, if carried out, will produce a value-increasing exchange fails . . . [and] nonenforcement may discourage such conduct in the future." Considering the case of withholding of information he argues, "It would be inefficient to require the seller of a house to obtain and disclose every fact that might be material to a purchaser; it would be particularly inefficient to require the seller to obtain and disclose information that the buyer could obtain at lower cost [footnote omitted]."[16] He does not consider the tougher case of nondisclosure of already obtained information of value to the purchaser. Efficiency at that point would call for disclosure. This would be a disincentive to the seller to obtain information which he might make use of but which might later negatively affect the selling price. It would also be a disincentive for the buyer to obtain information which he might presume the seller has checked for him, although this problem is easily handled by asking. This question is also considered separately (in §3.6), where Posner considers information flows between seller and consumer, arguing that information about simple, frequently purchased,

---

here is consideration of the total set of costs, not just those of the single transaction.

[13] Obviously legal placement of liability independent of the contract does not permit this shifting.

[14] Posner also considers the reverse case of compelling market rather than legal exchanges where the former are cheaper, as in determination of price, when the parties have been vague.

[15] P. 44.

[16] Pp. 48–49.

products subject to patent defects is most cheaply gathered by the consumer; while information about complex infrequently purchased products subject to latent defects will be cheaper for the seller to obtain. Posner also considers other types of information flows in the market, problems of suits for damages where damages are small and Federal Trade Commission's regulation of advertising.

The implicit background for this discussion is a conventional resource allocation model where markets are functioning more or less properly, so one can focus on the single transaction against a background of efficient resource allocation. But it is natural to ask whether the same information and transaction costs which create the questions at hand do not simultaneously imply some systematic biases in the entire market setting which are relevant for the treatment of the single transaction and efficiency. As an example, let us consider Posner's discussion of standard form contracts (§3.7). Many offers are made on a take-it-or-leave-it basis by individual firms. This pattern may well reflect significant cost savings for a firm engaged in many small transactions. But, he argues, competition will adjust for this. "If one seller offers unattractive terms to a purchaser, a competing seller, desiring to obtain the sale for himself, will offer more attractive terms. The process should continue until the terms are optimal from the purchaser's standpoint. Thus the purchaser who is offered a printed contract on a take-it-or-leave-it basis does have a real choice: he can refuse to sign, knowing that if better terms are possible another seller will offer them to him. All of the firms in the industry may find it economical to use standard contracts and to refuse to negotiate with purchasers. But what is important is not whether there is haggling in every transaction but whether competition forces sellers to incorporate in their standard contracts terms that maximize the purchaser's benefits from transacting."[17]

But what has happened to all those information limitations and transaction costs that were the heart of the argument? For terms of limited importance is it worth the cost of seeking a better contract? (Can you be sure one even exists?) Is it even worth the cost of finding out what would be a better term?[18] If the answers to these questions are no, where is the incentive for better terms? Might not all firms find it to their interest to offer bad terms[19] (just as models have been constructed where prices exceed the competitive price because information limitations remove the advantages of price decreases)? If in some cases the market systematically diverges from efficiency, is not the role of a legal system's seeking efficiency basically different from that analyzed?[20] Perhaps no rules to cure this problem can be found which are better than those which ignore it. Perhaps the legal system can only help efficiency for the (excessively?) well informed. Apart from the obvious distributional questions, this appears to be an extremely important efficiency question, the outcome of which is not at all clear.

---

[17] P. 54.

[18] Examples would be risks better borne by the seller, so the consumer benefits from better terms and a higher price.

[19] Free entry cannot come to the rescue here, for even with identical firms we merely get zero profits, not efficient terms.

[20] These problems are further complicated by the presence of plain old-fashioned careless consumers in addition to new-fangled transaction-cost-avoiding careless consumers.

As I hope I have indicated by this discussion, Posner is examining an area of great intrinsic interest and analytical difficulty and making considerable progress. The book is broad as well as deep. I found it fascinating to read and I look forward to using it as a text.

Noll, Peck, and McGowan's *Economic Aspects of Television Regulation*

Reviewed by
Stanley M. Besen

Associate Professor of Economics
Rice University
and Consultant
The Rand Corporation

and

Bridger M. Mitchell

Senior Staff Economist
The Rand Corporation

1. Introduction

■ In *Economic Aspects of Television Regulation*,[1] Roger G. Noll, Merton J. Peck, and John J. McGowan, hereafter referred to as NPM, essay a comprehensive analysis of the regulation of the television broadcasting industry. To give some idea of the breadth of their efforts, a partial list of the topics covered includes the employment patterns of FCC Commissioners, the financing and organization of public broadcasting, the substitutability between television advertising and that in other media, the management of the radio frequency spectrum, the prospects for videocassettes, the cost of operating television stations, the anticompetitive effects of television advertising, the factors determining to whom a new television license will be awarded, the prospects for forming new television networks, the cost of television program production, and the future of subscription television. The reader who seeks a comprehensive introduction to the major issues concerning the regulation of television broadcasting can do no better than to read this book.

While *Economic Aspects of Television Regulation* will be an outstanding primer for newcomers to the analysis of the broadcasting industry, it asks to be much more. The book is intended as a statement of the authors' views on the major policy questions in broadcasting, and here it is less successful. Many of the judgments reached by NPM obtain no support from their own re-

---

Preparation of this paper was supported under a grant from the John and Mary R. Markle Foundation to the Rand Corporation. The authors have benefitted from discussions with L. L. Johnson and R. E. Park, from the friendly criticism of B. Ellickson and G. Nelson, and from the editorial assistance of A. J Coerver and M. Fisher. The views expressed here are those of the authors and do not necessarily represent those of the Markle Foundation or of The Rand Corporation.

[1] Washington, D. C., The Brookings Institution, 1973.

search or that of others. We say this with some disappointment because, while we are sympathetic with many of the authors' conclusions, an objective assessment would have to regard the cases for them as not proven. Rather than attempt the point-by-point defense of this assertion that would take us well beyond the limits of a review article, we have decided instead to examine in detail the authors' treatment of a few major issues. These include: the demand for television; the economics of network behavior; the financing of unremunerative services; and the analysis of regulatory behavior. In the final section of the paper we provide some speculations on the reasons for some of the book's shortcomings.

2. The demand for television

NPM devote considerable attention to the demand for television. Since the answers to so many policy questions depend upon an understanding of viewer demand, this represents a sensible allocation of resources. Both the strengths and the weaknesses of the book are vividly revealed in its treatment of this subject.

The analysis of viewer choice. In an early and influential analysis of radio broadcasting,[2] Peter Steiner assumed that individuals would listen only to their preferred program type and, if it were not available, would not listen. When combined with the assumption that stations seek to maximize audience size and with the "welfare" criterion that as many persons as possible should receive programs that they are willing to listen to, the Steiner assumption about listener preferences implies that competition in broadcasting may lead to "duplication" in programming fare. This is the case because a new station may find it more profitable to duplicate the programming of an existing station and share in its audience rather than to program for previously unserved potential listeners. Now because the persons who listen to the program of the new station were previously "satisfied" (since they were listeners before the new station came into existence), "welfare" is not increased when the second station duplicates the programming of the first.

Steiner's strong result stems principally from his assumption that people will listen only to programs of their first choice. When one allows listeners to have second choices, however, the analysis of listener (or viewer) behavior becomes much more difficult. One must now specify how choices are made among programs, none of which may be a viewer's first choice, as well as how different from his preferred program the available viewing alternatives must be for him to become a nonviewer. Moreover, the Steiner welfare criterion no longer suffices. People who are not receiving their first-choice programs will often be viewers nevertheless, and one must therefore specify the value of the viewing alternatives available to them.[3]

---

[2] See [10].

[3] Steiner's welfare criterion is not even appropriate, given his assumption about listener preferences, unless one also assumes that all viewers

A model of viewer behavior must describe how viewers value different programs. It must do so in order to determine which of a number of competing programs a viewer will watch and to compare viewer satisfaction produced under alternative market structures. Such a model should permit, for example, judgments about the impact of the introduction of pay television and about the effect of increasing the number of competing stations in a market.

A number of attempts have been made to improve on Steiner's approach. Wiles, for example, has developed a model in which he allows for differences in the value that viewers place on different programs.[4] While major work remains to extend the theory of viewer choice, some of the principal remaining difficulties in improving our understanding of viewer behavior are empirical. One of the most formidable problems is that it is difficult to describe the characteristics of programs in terms of which viewing choices are made. One can use industry program categories such as Westerns, situation comedies, variety, etc., but one cannot be sure that programs within these categories are closer substitutes for one another than they are for programs in other categories. Without information about the program characteristics that are important to viewers, judgments about viewer satisfaction that are made by comparing the availability of programs placed in industry categories may be seriously in error.[5]

But even if we knew which characteristics viewers regarded as important, an equally vexing problem would remain. Since almost all of the evidence that we have about viewing choices comes from situations in which television is provided "free" to viewers (the costs of programming being borne by advertisers), it is virtually impossible to measure the intensity of viewer demand. Moreover, there may be programs that might have obtained sufficient revenues to pay for themselves out of receipts from viewers but that could not attract large enough audiences to be shown over advertiser-supported television. But due to the tight restrictions on subscription television, market data on the demand for such programs are not available.

Faced with these difficulties, a number of strategies have emerged for obtaining *some* information about viewer preferences. One alternative is to examine the demand for cable television, since cable offers, at a price, alternatives not available to viewers over the air. Another approach is to analyze the evidence from the limited number of experiments with pay or subscription television. Finally, some have attempted to analyze the choices among programs when all programs command a zero money price. NPM have adopted each of these approaches at one point or another.

☐ **The evidence from "free" television.** The authors make the distinction between viewing *options*, the number of programs which

---

receive equal satisfaction from obtaining their preferred program type and that viewer satisfaction is additive.

[4] See [11].

[5] For an interesting attempt to identify the factors underlying viewer choices, see [2].

are simultaneously available to viewers, whether or not they are of the same program type, and *diversity*, the number of categories in which programs are offered at a given time. A central theme is that viewers value more options, i.e., the number of existing and utilized channels. Viewers apparently value more options for two reasons. First, the more channels that exist, the larger will be the number of differing program types. NPM cite evidence that diversity increases, although at a decreasing rate, as the number of stations in a market increases. Second, NPM recognize that programs within standard categories may not be perfect substitutes for another and that this may cause viewers to value more options even if diversity, as conventionally measured, does not increase.

The authors summarize their basic position on viewer choice in these words:

> Since the price of television watching is zero, . . viewers seek to watch each program type until its marginal benefit to them is also zero. Similarly, because of the zero price the networks have no incentive to broadcast programs that provide benefits to the viewers more than incrementally above zero, since they will collect no greater fee from those who are satisfied beyond the level necessary to induce them to watch the program in the first place This arrangement implies that a large number of viewers regard the benefits from an increment to viewing time in general as negligible, but regard the benefits of additional hours of certain types of programming as very great.[6]

Apart from the empirically unsupported references to "large number" of viewers and "very great" benefits, the analysis does not produce the asserted conclusion. The most important flaw is in the assertion that the networks have no incentive to provide benefits to viewers more than incrementally above zero. But, of course, competition among the networks will mean that they will be forced to do so in order to attract audiences. Moreover, even if there were only a single station, its programs would provide satisfaction in excess of zero to inframarginal viewers. Such a station would choose programs and expend resources on them to the point at which the cost of attracting an additional viewer would be just equal to the additional revenues that advertisers would pay for access to that viewer. While there may be viewers who would watch the program and receive little satisfaction from doing so, there would be others who would have watched and been willing to pay a positive price to do so.

A second difficulty with the above analysis is that it does not lead to the conclusion that "the benefits from an increment to viewing time in general [are] negligible." And, elsewhere in the book, NPM advance evidence that suggests that total viewing will increase as the number of viewing alternatives increases.

But, despite the deficiencies in the analysis, the authors' basic conclusion, that viewers value additional viewing alternatives, is almost certainly correct. All that is required to reach the conclusion is that there be differences in preferences among viewers. Given any limited number of viewing alternatives, some viewers (although the number need not be "large") would switch to other programs if they were available, because they place a lower value on the programs they are currently watching. The result depends neither on the "unwillingness" of the networks to provide programs valued only a little above zero nor on any constancy in total viewing. In fact, the conclusion could have been reached with some of the other models of viewer preferences that the authors discuss elsewhere in the book.

---

[6] [6], p. 28.

What evidence do the authors provide on the nature of viewer preferences? NPM cite a ten-year-old survey of New York City viewers as evidence that the "greatest unfilled demand is for just those types of programming that already dominate the network schedule."[7] In the survey, viewers were asked what types of programs they would like to have more of, and the first type mentioned was recorded. Aside from the obvious difficulty in ascertaining viewer preferences by means of such a question, the answers do not reveal whether viewers would substitute these programs for ones they are watching or whether they would expand total viewing. Nor do they tell us how much these additional viewing alternatives are valued. In short, the survey fails to provide precisely the kind of information that is required if one is to choose intelligently among alternative structures for the television industry.

Still another source of evidence that NPM advance to support their assertion that viewers want "more of the same" is a regression relating the share of the audience received by a network affiliate to the network with which it is affiliated and to the number of competing affiliates and independent stations it faces.[8] The results of the regression imply that the total audience share obtained when there is only one affiliate in a market is about 45 percent of the potential audience, that two affiliates together will receive about 55 percent, while with three affiliates the figure rises to about 60 percent. Further, each affiliate receives a nearly equal share of the audience. Aside from the fact that this evidence conflicts with the earlier statement that the aggregate amount of viewing is independent of what is shown, the evidence does not support the authors' statement that "viewers value more options as well as diversity."[9] What the evidence indicates is that competing stations divide equally an audience which grows as the number of competing stations increases. The expansion in total audience reveals that the program of the second station is somewhat different from that of the first because the new station succeeds in drawing its audience both from that of the existing station and from the ranks of nonviewers. But to argue with this evidence that viewers value a larger number of options requires the belief that the increase in the audience would have occurred even if the second station had provided programming much like that of the first.

NPM use the same evidence to argue that "a two channel monopolist would probably not find it profitable to behave much differently from existing networks."[10] But this argument is not supported by the evidence. Apparently what the authors had in mind was that the increase in total viewing as a result of the entry of the second station implies that it provides programming different from that of the first. But this does not mean that a single entity controlling more than one station would provide the same pair of programs. We would expect, instead, that it would select its programs so that if one program were eliminated, a minimum number of viewers would switch to watching the other program. In other words, we would expect programming even more differentiated than that of the two competing stations.

---

[7] [6], p. 9.
[8] [6], p. 51.
[9] [6], p. 51.
[10] [6], p. 52.

☐ **Subscription television.** The unsatisfactory treatment of viewer choice appears most starkly when NPM evaluate the prospects for subscription television. The success of subscription television depends on the extent to which viewers are willing to pay for programs that are different from those available on advertiser-supported television. The authors begin their analysis by arguing, correctly, "the responsiveness of audience size to price variations —the price elasticity—is the critical factor that would permit the intensity of demand to be measured against mere number of viewers."[11] They go on to argue that "price elasticities may indeed differ among programs. But even if they do, the networks are unlikely to set prices accordingly."[12] This leads them to conclude that "if all mass entertainment programming were offered at roughly the same price . . . networks would soon pick that mix of programs that maximized audience size". and "the only difference from the present system would be that the single price charged would be positive instead of zero."[13]

The argument used to support the view that prices would not differ among programs is both dubious and inconsistent with evidence from the pay television experiments reported by NPM. Even if it were correct, however, it does not follow that the programs that would maximize audience at any positive price are the same as those that would maximize audience at a price of zero.[14] So long as there are people who would become nonviewers as the price of television programs rises, the programming mix is likely to depend on the established price. Only under very special conditions will the programming mix be the same. The conclusion that "the only difference from the present system would be that the single price would be positive instead of zero" is simply not correct.

---

[11] [6], p. 131.

[12] [6], p. 131.

[13] [6], p. 132.

[14] One of the most annoying features of the book is that it is not always clear what the authors' view is on a subject despite the fact that it is discussed in a number of places. To take only one example, they argue (p. 132) that prices for subscription television programs would not reflect "the relative costs and popularity of the programs." But on p. 139 we are told that the specific program charges in the various pay television experiments depend on "an estimate of the program appeal." The argument "the best strategy might be to avoid price premiums on the popular programs" (p. 131) is inconsistent with the observation that program charges ranged from $0.25 to $1.50 (p. 139) with the higher prices being charged for live sports events, one of the most popular of the program types in the subscription television experiments.

NPM state that, if VHF channel capacity is limited, minority programs would be unlikely to outbid mass audience programs for time on VHF stations under pay television. Their argument is that a mass audience program attracting 15 million homes could, by charging a price of only 10 cents per program, outbid a minority program appealing to 1 million homes willing to pay $1.50 each for the time. Moreover, they argue that an audience of 1 million homes probably would not be attracted at a price of $1.50, and that if only half that audience were reached, a charge of $3 would be necessary.

But, of course, it is easy to turn NPM's example on its head. *If* 3 million homes could be attracted to the minority program, *then* a charge of only $0.50 would be sufficient to permit the minority program to compete with the mass audience program. The point is not that these numbers are any better than NPM's, but rather that without any evidence on that "critical factor," the price elasticity, neither they nor the reader can make any judgments about the ability of minority audience programs to bid for time against mass audience ones. The basic question is still: How many people value programming different from that available on free television and how much are they willing to pay for it? Without the answer to this question judgments about the future of pay television cannot be made. Comparing the actual audiences that free television programming attracts with hypothetical numbers of viewers and the prices that subscription television would obtain simply does not suffice.[15]

In fairness to NPM, however, they have examined the results of a subscription television experiment authorized by the FCC for Hartford, Connecticut, and they merit credit for using these data for the first time. Unfortunately, their efforts are not carried out in a systematic and careful way. For example, they argue that the data indicate that about 4 percent of all homes would pay $12 per month for subscription television when they already have available three network stations, at least four independent stations, and public broadcasting,[16] whereas in Hartford 4 percent of households subscribed when there were *no* independent stations while the experiment was in progress. The reader is frequently provided with hypothetical numbers instead of real ones—"a more likely result would be that virtually all of the cable subscribers would pay $1 a month for the extra service."[17] A more important shortcoming is that NPM use observations on the audiences for various program types to claim that "the strongest unfilled consumer demand is in categories similar to those that dominate network schedules."[18] While the observation itself is open to question—high school basketball and "concerts, opera, and ballet," both of which attracted over 10 percent of all pay television subscribers, are scarcely categories that "dominate current network schedules"; even if this were correct, it may reveal more about the limitations of the program categories used than about viewer preferences. After all, the category "movies," that has large audiences both on free television and in the Hartford experiment, contains both "Mary Poppins" and "Deep Throat." To formulate conclusions regarding what the Hartford experiment tells us about viewer preferences, something more is needed than a count of the number of subscribers viewing programs by broad program categories. NPM are to be commended for focusing our attention on the Hartford experiment as a potentially important and unexploited body of data, but the data still remain largely unexploited.

---

[15] The (relatively) unlimited number of cable channels available implies that minority subscription programs need not obtain an audience large enough to be as profitable as mass audience ones, but only large enough to cover their costs. This fact and the lower hardware costs of denying access to nonpaying viewers increase the likelihood that subscription television will develop in the cable rather than the broadcast mode.

[16] [6], pp. 147–148.

[17] [6], p. 148.

☐ **Cable and the value of free television.** The most widely used evidence on the demand for television arises from experiences with cable television. Cable, carrying almost exclusively existing advertiser-supported programming and thus far available primarily in areas with few local signals and/or reception difficulties, is a natural data source for analyzing the value that consumers place on additional viewing alternatives. In addition to frequent references to measures of consumer surplus from free television based on evidence from the demand for cable, NPM devote two chapters and an extensive appendix to an examination of cable television and its future.

In Appendix A, the authors begin by setting a high standard of econometric work. They construct a model of cable television demand using a Cobb-Douglas utility function in "quality of television service" and expenditures on other goods. Variation in tastes among consumers is captured by an exponential distribution of the ratio of the elasticity parameters of the utility function. From these premises, NPM derive the estimating equation:[19]

$$\frac{\ln(1 - P/Y)}{\ln PEN} = \alpha + \Sigma\, \beta_i \ln X_i + \Sigma\, \gamma_j D_j,$$

where P is the annual subscription price, Y is household income, and PEN is penetration, the fraction of households who subscribe to cable in a mature system. The proxies for quality of service are X_i (comparative service variables for network, duplicate network, and independent stations that measure the number of cable signals as compared to signals available off-the-air) and D_j (dummy variables indicating the presence of time-weather, local origination, and educational channels).

The equation is estimated using data from a 1969 sample of 31 cable systems of 10,000 or more subscribers. The authors justify deflating all variables by the square root of the age of the system on the grounds that younger systems in the sample may not yet have reached equilibrium and that age is a proxy for the quality of over-the-air service.

---

[18] [6]. p. 142.

[19] [6]. pp. 283–284.

The results of estimating the model are broadly consistent with earlier studies of cable demand.[20] In areas where homes are unable to receive all three networks, over half the homes will pay $5 per month to receive network fare alone. Importing independent stations induces additional, but much smaller, gains in subscribers. NPM argue convincingly that this sample provides a direct measure of the value to viewers of "free" television, and strong evidence that the public desires an expanded number of network and independent station alternatives. Their numerical estimates of this surplus, which they caution are subject to some uncertainty, are surprisingly large and suggest that the average household would be willing to pay more than 5 percent of its income rather than do without three-network television service.

Because the major potential for cable television lies in the large urban areas of the country which are, as yet, almost entirely unwired, predictions of the demand for cable service in major markets are central to an analysis of the national prospects for cable, but not readily extrapolated from the demand observed to date. The authors appropriately observe that both their own estimates and those of two earlier studies are not representative of conditions in the large urban markets, almost all of which have available at least three network stations of good technical quality. They therefore obtain a second sample of 40 cable systems located in areas that enjoy good quality broadcast reception of all three networks. NPM consider these data as "highly representative," despite a response rate of only 10 percent to their mail questionnaire. When fitted to these data, the consumer decision model yielded unsatisfactory results. There is a negative effect of additional independent signals, a much smaller impact of more network signals than in the other sample, and little explanatory power. In particular, the value of free television implied by the results using the major market data is strikingly smaller than that implied by the results using the first sample.

At this point the authors remark that "other factors such as price, income, and alternative available entertainment options"[21] may explain differences in penetration rates. They therefore discard their initial analysis along with their model of viewer behavior in favor of what we may call Model II, a linear regression of the penetration rate on price, income, a single comparative service variable, a dummy variable for a "top 100 market," and the age of the system. According to NPM, the principal results of Model II are that the profit-maximizing price is not significantly different from the typical $5 per month, and that cable systems approach an equilibrium penetration quite rapidly.[22]

Can there be a reconciliation of the two models? The reader searches the text in vain. Model I implies price and income elasticities equal in absolute value and of the order of 0.75 to 0.9 for a system that imports four distant signals. In contrast, Model II has a price elasticity of —1.95 at the mean and implies an income elasticity of about 1.0.[23] The role of the age variable is also unclear. In Model I, age causes system disturbances to have unequal variances but no effect on the number of subscribers, whereas in Model II it has a positive effect on expected penetration.

---

[20] See, e.g., Comanor and Mitchell [3] and Park [7].

[21] [6], p. 296.

[22] [6], p. 156.

Indeed, Model II is nothing more or less than another of the "ad hoc linear relationships between cable penetration on the one hand and price, income, and quality variables on the other"[24] for which NPM have criticized other researchers. If these variables are valid determinants of penetration, an appropriate methodology would have been to incorporate them into a more fully specified Model I.[25]

Instead, NPM ignore the inconsistencies of their two models and proceed to use only selected results from each. Having acknowledged the need for data on systems operating under conditions similar to those in large cities and then having estimated Model II, the authors employ the results of their *first* model to predict the demand for cable in large cities. According to NPM, penetration will reach 50 to 70 percent in major markets if cable carries four independent signals. Under these conditions, cable will have "enormously large potential profits."[26]

Not only is this forecast based on data that are not typical of large city conditions, but the predictions are substantially at variance with those of the one previous study using data for systems operating in environments like those in the major markets. R. E. Park, in research published in this *Journal* in 1972,[27] analyzed data from a sample of 63 cable systems, all of which are in markets having at least three high-quality signals available off the air. For each system, data were verified in telephone interviews. Park's specification takes into account two important factors neglected by NPM —the variation in technical quality of off-the-air signals, and what has come to be called the "handicap" that UHF stations face in competing with VHF stations for off-the-air audiences. The resulting model is estimated by nonlinear methods. Park's work implies that penetration for urban systems will range from 20 to 45 percent.

---

[23] NPM report only that penetration increases by 0.05 per $1000 of income in 1972 prices. Means of 0.50 for penetration and $10,000 for family income would yield unitary elasticity.

[24] [6], p. 290.

[25] This is not the only case in which the authors' specification leaves something to be desired. Other instances are in their treatment of the demand for advertising (p. 41), in their analysis of the prices of television programs (pp. 78–79), and in their examination of the factors underlying a television license award (pp. 112–115). In the last case, for example, to measure the criteria actually used by the FCC to award licenses, NPM carry out a statistical analysis of 45 applicants for 16 licenses awarded in the 1967–1970 period. The dependent variable took on the value one if a license was granted, and zero if it was not. While the authors are concerned in several other regressions about the presence of heteroscedasticity, it goes unnoticed here. A more serious error is the failure to note that 16 observations are necessarily dependent, reducing the degrees of freedom available, and raising questions about the bias of the estimated coefficients.

[26] [6], p. 159.

[27] See Park [7].

Cable, according to NPM, would achieve 66-percent penetration of all homes offered service and would actually wire 54 percent of all U. S. homes *if* free to import four high-quality independent signals.[28] Such systems would be highly profitable, generating an above normal profit of $500 million and a consumers' surplus of $1.7 billion annually.

But this development will not be permitted. Instead, the FCC has severely restricted the number and types of signals that may be imported and has imposed numerous requirements that cable systems provide unremunerative services. The combined effect of these restrictions "will probably foreclose cable development in most large cities" (p. 204). Thus, although their estimates of the losses to consumers caused by these policies is greater, NPM foresee much the same prospects for cable under the FCC rules that have been predicted by Comanor and Mitchell.

3. The economics of networks

■ The three national television networks occupy a central position in the American television industry. Together with their affiliated stations they serve almost all homes in the United States. Most programming produced for television is initially sold directly to them. Most network affiliates are profitable, and most independent stations are not. It is therefore not surprising that NPM have devoted a good deal of attention to the economics of the networks.

☐ **A model of network behavior.** Television networks operate as intermediaries between their affiliated stations and the suppliers of programs. The networks bear the costs of program acquisition and of system interconnection. In return, they receive a share of the advertising revenues generated. The remainder of the revenues are paid to the stations for the use of their time.

In Appendix B, the results of which are employed in various chapters in the book, NPM develop a formal model of the relationships between a network, its affiliated stations, and program suppliers. The authors begin by observing "the share of revenue that must be given up to affiliates [in order to induce them to clear the full network lineup] will be minimized . . . if networks monopolize broadcast rights for the most popular programs. . . . Thus, the major means by which a network can hold down the share of revenues that it must yield to affiliates is the superior audience appeal of its programming relative to that available from nonnetwork sources."[29] Further, "the maximum share a network would have to offer a program supplier increases as audience size increases but decreases with the share of program revenues that networks give to affiliates."[30] Thus, the two elements in the network's decision making are "the need to provide an economic incentive to carry network programming establishes a lower limit on the share of revenue that networks must yield to affiliates" and

---

[28] Because of low housing density, low incomes, or high construction costs, some areas may never be wired.

[29] [6], p. 304.

[30] [6], p. 305.

"a network must be willing to pay program producers at least the opportunity cost of resources used in program production. Consequently, it would be fruitless for the networks to set (the maximum share paid to affiliates) so high as to make it impossible for a program producer to recover these costs."[31]

The problem that a network faces is to set the share of revenues paid to its affiliates on various programs high enough to get them to accept the programs and to pay program suppliers enough to induce them to supply their programs to the network.[32] The two payments clearly interact. The higher the payment to program suppliers, the better the programs attracted to the network and the smaller the share paid to affiliated stations; but the higher the payment to affiliates, the smaller the necessary payment to suppliers, because they will face a more difficult time in dealing directly with the stations. The analysis of the interaction between these two decisions represents an important insight.

But having established the linkage of these two decisions, NPM proceed to ignore their interaction. They assert "the networks would like . . . to set [the program owner's share in broadcast revenue] so that . . . program costs would . . . be minimized and profits maximized."[33] And they state "in view of the role of [the maximum sharing ratio for affiliates] in determining the share of broadcast revenues that established programs can extract . . . one would expect [the maximum sharing ratio]" to be set quite closely to "the value that would minimize the necessary payment to program suppliers."[34] But the networks do not wish to minimize the payment to program suppliers any more than they "have a preference for a high value of the maximum sharing ratio for affiliates."[35] They simply wish to maximize their own profits, and this requires neither a maximization of the payment to their affiliates nor a minimization of their payment to program suppliers. In fact, NPM find that payments to program suppliers exceed their resource costs.

☐ **The economics of program supply.** NPM use the observation that, on the average, program suppliers earn returns in excess of normal profits to argue throughout the book that additional resources available to television will have little or no effect on programming volume. But this conclusion results from a confusion between the marginal and the average supplier. They argue "further increases in the expected rate of return are unlikely to draw additional resources into the [program production] industry."[36]

---

[31] [6], p. 305.

[32] NPM argue that it will be difficult for the networks to implement a compensation structure for affiliates that provides differential compensation according to the popularity of network programming relative to nonnetwork alternatives. But existing compensation arrangements have precisely that character. See Besen and Soligo [1] for a fuller discussion of the network-affiliate relationship which deals explicitly with this question and with the regulatory environment in which the relationship exists.

[33] [6], p. 307.

[34] [6], p. 309.

[35] [6], p. 309.

[36] [6], p. 84.

But so long as there are variations in ability among program suppliers, it is not surprising that some suppliers earn differential rents, leading to average earnings for the industry that exceed returns expected at the margin. This no more permits the conclusion that the supply of programming would be unaffected by an increase in the return to programming than observing that, on the average, the owners of corn land earn rents would allow us to conclude that the amount of land devoted to growing corn is independent of the return to corn farming. One does not need the contrived argument that there is "a scarcity of ideas and talent for developing more pilots that have a chance of becoming successful network series"[37] to explain above-normal profits. Conventional economic theory will suffice.

☐ **The prospects for new networks.** One of the most widely heralded of NPM's policy conclusions is that, if channel allocations were changed to permit it, the nation could support six over-the-air television networks.[38] Given the extensive treatment of the behavior of the existing broadcast networks and their relationships with program suppliers and affiliated stations, one might expect that this important conclusion follows directly from that analysis. On the contrary, the conclusion is obtained from a simple extrapolation of costs and revenues based on data for the three existing networks. The thrust of the argument is that the programs being produced for the major market's independent station could be widely distributed at little additional cost and that the costs of the new networks based on the programming of these stations would be covered by the new revenues generated.

But no data are ever presented on the profitability of the three new networks. Instead, we are provided with estimates of the aggregate profits in a six-network system. None of the analysis of the relationship between the networks and program suppliers is used except to argue that, since industry revenues decline due to the elimination of local advertising, program suppliers' rents will fall. Since programming costs are approximately one-half of total revenues for the three existing networks, that proportion is used for the six-network system. Aggregate audience measures are derived solely from audience data for New York City. Revenues are obtained by extrapolating revenues per viewer from the three-network system. The model of the relationship between networks and their affiliates is never used to examine what would happen when the number of stations in each market is equal to the number of networks, nor is there any analysis of how the three existing networks might change their programming policies in response to the entry of the three new networks. In short, there is no analysis of the behavior of the participants in the industry at all. An important opportunity has been missed: an analytical model is developed to explain behavior, but when an important problem to which the model can be applied is addressed, the analysis is jettisoned in favor of an *ad hoc* analysis. As a result, one does not know very much more about the prospects for new networks after reading the book than he did before.

---

[37] [6], p. 84.

[38] [6], pp. 117–118.

4. Financing unremunerative services

■ Regulated industries abound in hidden subsidies, and the television industry is no exception. NPM identify a number of such subsidies, attempt to measure their costs and benefits, and seek more efficient means of achieving the same ends.

☐ **Localism.** The authors are at their best in illuminating the long-standing and perpetually ineffective FCC policy of localism, the promotion of "local service" broadcasting by establishing stations in as many localities as possible. This goal, which was first articulated in the early days of radio, became an important consideration in developing the Commission's master plan for allocating television broadcast frequencies. In promoting localism, the FCC explicitly rejected the Dumont proposal for a television system based on powerful regional VHF stations in order that "as many communities as possible should have the opportunity of enjoying the advantages that derive from having local outlets that will be responsive to local needs."[39]

The Commission has sought to expand the range of viewing alternatives by allocating a large portion of the frequency spectrum in the Ultra High Frequency band to television. But even though almost all television receivers currently can receive UHF television, UHF stations without network affiliations are almost all unprofitable, as are many network-affiliated UHF stations. The limitation on VHF channels in many markets has restricted the number of viable networks to three. Thus the policy of localism combined with the handicaps that UHF stations face has produced a situation in which many viewers have fewer viewing alternatives than are technically and economically feasible.

Given the well-known costs of the policy of localism, what are its benefits? NPM's survey of Commission hearings and investigations concludes that, aside from local news and sports which are generally self-supporting, "local service is a myth."[40] The judgment is plausible, both because local programming is costly to produce and because it is of limited interest to viewers who have the alternative of watching network or nationally syndicated programming. Nevertheless, NPM themselves might have investigated the local service programming of various stations in order to give more support to their own conclusion.

☐ **The cable surplus.** NPM pointed out that the notion of a cable television "surplus"—the above normal profits that would be generated by unregulated cable development and can be applied to other uses—is a difficult one to resist, and this remark characterizes their own pronouncements on the regulation of cable television as well as the FCC's. In an earlier paper that they prepared for the

---

[39] Quoted in [6], p. 101.

[40] [6], p. 110.

Sloan Commission of Cable Communications,[41] the authors entertained three categories of "realistic proposals." It might be necessary to pay part of the surplus in the form of "bribes" to various interest groups in order to persuade them to permit cable development to occur. While not necessarily preferred by NPM, the proposed recipients included the FCC (establishment of a UHF station development fund), network affiliates (compensation for reduction in profits), local governments (cable franchise fees), and program producers (copyright payments from cable systems). A second category of proposed uses of the surplus is reiterated in the book. The authors consider a host of unremunerative services to be "worthy claimants," including the FCC's earlier proposal of a "public dividend" tax on cable revenues for support of public broadcasting, free public broadcasting and free instructional cable channels, reduced subscriber rates in poverty areas, and mandatory live origination in major markets. And following this list of special interests NPM identify a third category for the "final claimant"—the consumer.

The reader senses the authors' recurring ambivalence in discussing proposals for cross-subsidization. How strong is the case for requiring the production of any of these special services? Posner has proposed that the burden of proof should be on those who advocate such "taxation by regulation."[42] The efficiency losses of permitting cable systems to earn excess profits in order to subsidize the regulator's planning objectives can be particularly large.[43] This is especially true if cable subscription demand is price elastic, and the authors appear to believe that it is,[44] although this conflicts with both their first model of cable demand and the assumption used in their earlier Sloan Commission paper.

In concluding their discussion, NPM do criticize the FCC for imposing, through its cable regulations, "a high indirect tax on cable television viewers, partly in the form of higher subscription prices to cover the costs of unnecessary services and partly by limiting the access to viewers to services they value."[45] Yet rather than resisting the temptation to use the surplus, NPM have themselves entertained a wide spectrum of special subsidies, and in doing so have perhaps encouraged regulators to engage in this device of "planning by regulation."

☐ **Public television.** NPM devote an entire chapter to the financing and organization of public broadcasting. They conclude a discussion of the rationale for public television by arguing that, given growth in cable and subscription television, only a very limited case can be made for it.[46] Further, since the principal beneficiaries of public television are its viewers, "a method of financing that relates the available funds to viewer satisfaction is preferable to automatic, periodic payments from a widely levied tax."[47] Given the above, the reader is surprised to read several pages later that the authors' policy recommendation is that public television be provided with "automatic general revenue funds by formula with virtually automatic appropriations."[48]

---

[41] [5].

[42] See Posner [8].

[43] For an analysis of this issue see Comanor and Mitchell [4].

[44] [6], p. 298.

[45] [6], p. 205.

[46] [6], p. 216–217.

Little has intervened between pages 219 and 234 to prepare the reader for this startling reversal, and he is not told why it has occurred. Is it that NPM believe the developments in cable and subscription television that would render public television superfluous will not be permitted to occur? And, if so, why do they not then favor a financing method linking funding to viewer satisfaction? It is not until 10 more pages have passed that the mystery is resolved.[49] The authors propose that public television should be financed largely by means of Federal matching grants based on the amounts of funds it receives in the form of individual contributions. But the rationale for this proposal, both in the amount of funding it envisions and in the financing method, is not clear. The matching feature does provide for a reflection of viewer preferences, but the authors have argued previously that benefits to nonviewers are meager. The funding mechanism would place substantial costs on nonviewers and, more importantly, would make permanent a public television system for which the evidence, the authors conclude, consists only of "lonely shreds."[50] Although not an isolated instance, this may well be the most striking case of the book's separation of analysis and policy recommendations.

NPM consider an especially novel approach to financing public broadcasting—that public television offer some programs on a subscription basis, with the profits being used to subsidize the costs of free public television.[51] They argue that it is likely that public subscription television would be subject to fewer restrictions than commercial subscription television because there would be a smaller fear—presumably on the part of commercial broadcasters—that it would siphon programs from free television. But NPM never discuss the motivations of the public television operator.

Whatever the case for public television, it surely must rest on an analysis of what public television stations would provide under alternative regulatory environments and financing methods. If subscription television were permitted on public broadcasting, one would like to know, for example, what programs a public television operator would show when he is charging for television, what prices he would charge, and which programs he would subsidize. When discussing the behavior of FCC Commissioners (see below), the authors have attempted to explain their behavior in terms of underlying economic motivations. One can apply to public broadcasting, with appropriate modifications, a statement made by NPM in that context: "What is needed is at least the beginnings of a formal theory of the [public broadcasting station] so that the positions taken by [public broadcasters] can be better understood and perhaps even predicted."[52] But no such analysis appears in their consideration of public broadcasting, and without knowing what it is that public broadcasters wish to "maximize," an evaluation of public broadcasting and of alternative methods of financing it cannot be complete.

---

[47] [6], p. 219.
[48] [6], p. 234.
[49] [6], p. 244.
[50] [6], p. 217.
[51] [6], pp. 232–234.

5. Regulatory behavior and policy change

■ Since the authors have written a normative book and hold strong views for modifying the present regulation of television, it is of particular interest to examine their analysis of FCC behavior in conjunction with their own policy recommendations. Their view of the Commission is built on the premise that the Commission's decisions result from rational, optimizing behavior and that potential for economic gain motivates political as well as economic activity.[53] Other things equal, several characteristics of the regulatory environment lead regulators to prefer to minimize visibility and change. The authors predict that the Commission will be most receptive to groups with a high per capita interest in a decision, that it will tend to seek compromises, parceling out some gains to all interests in the name of fairness; and, when faced with an unavoidable appeal to outside judgment, it will delay a decision and attempt to shift responsibility to some other authority (the Congress, the courts, etc.).[54]

How do the authors see changes occurring? As advocates of consumer sovereignty they are acutely aware that consumers, with a small per capita interest, have not been well represented before the Commission. And they have observed that "reforms in the structure of broadcasting are far more likely if some protection can be afforded to existing interests and if some other existing, powerful special interest can be made to regard its own welfare as coincident with the public interest in the change."[55]

NPM's own position is that "a new legislative mandate for broadcast regulation is necessary."[56] In its principal features this mandate would divest the FCC of control of program quality and content, establish competition as the primary mechanism for reflecting viewer preferences, downgrade localism, and limit the license renewal process to consideration of engineering questions. They would also rely on expanded access to the airwaves to replace consideration of "fairness doctrine" and "equal time" questions.

The difficulty with this prescription is that legislative reform is no easier to bring about than it would be to get the Commission to change its behavior. The same special interests who have influence with the Commission are at home, as well, in the halls of Congress. The Commission can make few important decisions without Congressional approval, and criticism of the Commission is, therefore, criticism of the Congress. The recent case of the cable television "consensus," where fear of criticism from the Congress, in response to anticipated pressure from broadcasters, led the Commission to limit cable development more severely in February 1972 than it had proposed the previous August, is an outstanding example of this phenomenon, and one cited by the authors.[57]

---

[52] [6], p 120.
[53] [6], pp. 120–121.
[54] [6], pp. 124–126.
[55] [6], p. 128.
[56] [6], p. 275.

6. Conclusion

■ In the final chapter of *Economic Aspects of Television Regulation*, the reader is treated to an intelligent assessment of many of the major issues in the regulation of the television industry. The chapter represents the kind of mature statement about public policy that one would expect from its authors, and we suspect that many of the judgments they reach are correct. Our attention is once again focused on the frustration of the public interest accomplished by a regulatory agency. It is no matter that consumers prefer more television programming and are willing to pay for its delivery by cable or scrambled signal; the Federal Communications Commission is determined to separate the viewer from any real measure of consumer sovereignty in order to protect the wealth of broadcast station owners. Not content to establish fair "rules of the game" and to leave the outcome to competitive forces, the Commission finds itself regulating the behavior of the participants in the industry in ever increasing detail. The authors' call for substantial deregulation of television, for unfettered competition from new communications technologies, for the abandonment of the policy of localism in order to expand the range of viewing choice, and for divesting the Commission from control over program content is, therefore, welcome.

But many, if not most, of the conclusions reached could have been obtained independently from the analysis carried out in the other 325 pages. Perhaps the authors have recognized that policy judgments do not wait for definitive answers from the research community, and decided to advocate their position even though its analytic foundation is not yet sound. And here it is difficult to fault the authors. In their emphasis on greater reliance on the market and in the wide attention their conclusions will attract, the publication of this book will have a salutary effect.

Yet the authors have sought to give the impression that their conclusions are derived from a comprehensive analysis of the television industry. But no such analysis existed before they began their work and they have failed to provide one. Study of the broadcasting industry has, until quite recently, attracted only a minimal fraction of the economics profession's resources, and progress on both theoretical and empirical questions has been proportionately slow. Given the level of understanding at the time the authors began their work, it would be surprising if a definitive analysis of the industry emerged from their book.[58]

Modest and qualified assessments are the best the economist can offer at this time. In an attempt to buttress their position in the advocacy process, however, the authors have exaggerated the conclusiveness of their analysis and claimed far more support for their policy judgments than is warranted. Contrary to its appearances, *Economic Aspects of Television Regulation* leaves many, if not most, of the important questions concerning the television industry still unanswered.

---

[57] [6], p. 206.

[58] However, as we have pointed out above, we authors have not always taken advantage of existing work and, moreover, have committed a number of avoidable errors.

References

1. BESEN, S. M. AND SOLIGO, R. "The Economics of the Network-Affiliate Relationship in the Television Broadcasting Industry." *The American Economic Review*, Vol. 63, No. 3 (June 1973), pp. 259–268.
2. BOWMAN, G. W. AND FARLEY, J. U. "TV Viewing: Applications of a Formal Choice Model." *Applied Economics*, Vol. 4, No. 4 (December 1972), pp. 245–259.
3. COMANOR, W. S. AND MITCHELL, B. M. "Cable Television and the Impact of Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 154–212.
4. ———. "The Costs of Planning: The FCC and Cable Television." *The Journal of Law and Economics*, Vol. 15, No. 1 (April 1972), pp. 177–206.
5. MCGOWAN, J. J., NOLL, R. G., AND PECK, M. J. "Prospects and Policies for CATV" This was the *first draft* of a paper prepared for the Sloan Foundation in March 1971. The *final revised version*, published in October 1971, is available upon request from the Sloan Foundation in New York. An abridged text appears as Appendix B in [9].
6. NOLL, R. G., PECK, M. J., AND MCGOWAN, J. J. *Economic Aspects of Television Regulation*. Washington, D. C.: The Brookings Institution, 1973.
7. PARK, R. E. "Prospects for Cable in the 100 Largest Television Markets." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 130–150.
8. POSNER, R. A. "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.
9. SLOAN COMMISSION ON CABLE COMMUNICATIONS. *On the Cable: The Television of Abundance*. New York: McGraw-Hill, 1971.
10. STEINER, P. O. "Program Patterns and Preferences and the Workability of Competition in Radio Broadcasting." *Quarterly Journal of Economics*, Vol. 66, No. 2 (May 1952), pp. 194–223.
11. WILES, P. "Pilkington and the Theory of Value." *Economic Journal*, Vol. 73 (June 1963), pp. 183–200.

Approving nuclear power plants: scientific decisionmaking or administrative charade?*

Paul L. Joskow

Assistant Professor of Economics
Massachusetts Institute of Technology

1. Introduction

■ The past decade has witnessed a dramatic increase in the nation's concern over the environmental effects of economic growth and technological change. These concerns have been given a legal basis in various local, state, and federal statutes aimed at protecting the environment. The basic federal environmental policy has been expressed in the National Environmental Policy Act (NEPA) which went into effect on January 1, 1970. NEPA requires that an environmental impact statement must be filed along with all actions of the federal government and associated administrative agencies which affect the human environment. NEPA does not, however, define exactly what form these impact statements should take, nor how they should be evaluated.

Under the scope of the act come the licensing and approval activites of the Atomic Energy Commission (AEC) and the Federal Power Commission, with regard to electric generating and transmission facilities. In 1971, Circuit Court Judge Skelly Wright brought the AEC's licensing approval process for nuclear generating facilities to a complete halt.[1] Judge Wright held that the licensing procedures used by the AEC for approving the construction and activation of nuclear generating facilities did not conform to the intent of Congress in NEPA. He held further that the AEC was required to draw up a detailed scheme for ascertaining the costs and benefits associated with proposed nuclear facilities, including both private and social benefits and costs, and to weigh the benefits against the costs in determining whether approval would be given to the facility in question. In essence, Judge Wright was requiring the AEC to perform a cost-benefit analysis of all proposed nuclear generating plants and to approve only those where the benefits outweighed the costs.

The AEC has responded to the dictates of NEPA as interpreted by Judge Wright through Regulatory Guide 4.2, *Preparation of Environmental Reports for Nuclear Power Plants*, issued March 2, 1973. This Guide seeks to lay out in detail the format and content of acceptable environmental reports, which are now required to accompany all applications for construction and operating permits for nuclear power plants.[2] This review will discuss the methodology prescribed by this Guide in the context of the environmental problems associated with nuclear power, microeconomic theory relating to externalities, and particular characteristics of private firms operating within the public utility regulatory process. To facilitate discussion, the review is composed of four parts. Section 2 discusses the environmental problems associated with nuclear power plants. Section 3 of the review presents the approach that the AEC has taken for dealing with these problems through Regulatory Guide 4.2. Section 4 critically discusses the cost-benefit analysis required by the guide, both in terms of its theoretical validity and its likely effects. Finally, Section 5 suggests that nuclear power may present special problems which vitiate the practical usefulness of traditional approaches to externality problems.

---

* A review of the AEC's *Preparation of Environmental Reports for Nuclear Power Plants* [2].

[1] See [1].

[2] The AEC then uses the information provided in this report along with its own independent analysis to prepare an environmental report which accompanies approvals of all applications.

2. Environmental problems associated with nuclear power plants

■ Since the secrets of nuclear power were first uncovered, many scientists have sought to harness this power for peaceful purposes. The first nuclear power plants were constructed in the mid-1950s, but by the end of 1973 there was a total of only 42 operable nuclear power plants—about 5 percent of total electric utility generating capacity. This small number should not mislead us into thinking that nuclear power is a minor element in the future of electricity supply, however. On the contrary, today there are over 140 nuclear generating plants being built or in advanced planning stages. While estimates differ, current plants call for about 25 percent of electric generating capacity to consist of nuclear plants by 1980 and perhaps 50 percent nuclear by the year 2000. At current rates of growth in electricity demand this implies the construction of as many as 1,000 nuclear generating plants by the end of the century. There is no question that a major commitment has been made by the electric utility industry to provide a substantial portion of additional power requirements through nuclear generating facilities.

A great deal of public concern has been directed toward the use of nuclear power generation. This concern has been directed at three types of environmental and safety problems associated with nuclear generating plants: (1) waste heat; (2) emission into the biosphere of radioactive substances (either by design or by accident); (3) disposal of radioactive wastes.

The waste heat problem arises from the use of water to cool down the steam in the condenser. Cooling water leaving the condenser will be about 20°F warmer than when it entered. This excess heat may be discharged directly into the cooling water source (lake, river, etc.) as warm water or as warm, dry, or wet air through the use of cooling towers or cooling ponds. For various reasons light water reactors discharge significantly more heat to the cooling water than do comparable fossil-fuel plants or high-temperature gas-cooled reactors which have a higher thermal efficiency and discharge part of their heat through stack gases. Environmental concern regarding reactor cooling water has been directed toward the effects on the ecology of the cooling water source—of both rising water temperatures and the ingestion of large amounts of water into the reactor cooling system.

Nuclear power plants discharge minute qualities of radioactive substances into the cooling water and directly into the atmosphere in the course of normal operation. The levels of such normal discharges have been designed to be kept within limits set by various government agencies. Additional radioactivity can leak into the environment as a result of minor failures in storage and piping systems, and thus allow small additional amounts of radioactive material to enter the cooling water or the atmosphere. More severe leakages of radioactivity may occur as a result of accidents during refueling or a major failure in the reactor core, including the failure of the reactor cooling system and emergency backups, causing the leakage of primary fissionable materials from the reactor vessel.[3] Extensive safety systems have been designed to make the probability of such major system failures, leading to the release into the atmosphere of large and dangerous amounts of radioactive material, extremely small. Theoretically, the levels of natural radioactive emissions and the probability of major accidental discharges can be reduced to arbitrarily small, but still positive, values by spending more and more money on containment and safety devices.

Environmentalists have expressed concern about radioactive emissions on two counts. They are concerned about the natural emission levels that have been approved, because they believe that we really do not have enough experience to know whether the so-called safety limits are really safe. Second, they believe that the reliability of the safety systems designed to mitigate accidental radioactive discharges may be very low and that increasing numbers of even larger reactors, which rely on computer simulation models or extrapolations of limited experience from smaller reactors, are being built. In short, they believe that the probability of a serious radioactive discharge is actually much higher than designers indicate. Finally, there are some who feel that however small the probability of a serious accident, the consequences of such accidents could be so catastrophic as to eliminate effectively nuclear power as a viable energy supply alternative.[4]

Nuclear reactors must be periodically refueled as the fissionable material is spent. After a suitable cooling period the fuel cores are sealed in shielded containers and are transported to a reprocessing plant. At the reprocessing plant residual uranium and plutonium are recovered, leaving a highly radioactive liquid residue which must be stored in liquid form for about five years and then be evaporated to leave solid radioactive residues. These solid residues are then to be shipped to a depository for an indefinite storage period.

Concern about the radioactive waste problem has been growing. Critics point to the fact that by the end of the century hundreds of tons of radioactive waste will be shipped over the nation's transport system each year. They believe that the potential for transport accidents is high and that the results of such accidents (with shipments contemplated as large as three tons) could be severe. In addition, these critics argue that no satisfactory solution to the long-term storage problems, associated with ever larger amounts of solid radioactive wastes, has been put forward.

---

[3] The leakage of primary fissionable materials into the ground under a nuclear plant has been referred to as the "China Syndrome."

[4] Estimates of the effects of a major accident differ tremendously. Estimates that I have seen of the results of a major accident involving a relatively small reactor range from 3,500 deaths and $7 billion in property damage to 45,000 deaths and $17 billion in property damage.

Proponents of nuclear power are of course aware of the problems that have been raised by environmentalists. They believe, however, that they already have dealt or will be able to deal with them effectively. While nobody would contend that nuclear reactors will lead to no damage to the environment, nor that the probability of a serious accident is zero, they emphasize that alternative energy sources also damage the environment and may result in the loss of human lives. All things considered, nuclear power appears to them to yield the greatest net benefits to society. It is this kind of implicit cost-benefit calculation which the AEC's Guide 4.2 tries to make explicit.

3. The AEC environmental guide

■ The Atomic Energy Commission's Guide 4.2 requires that each applicant submit an environmental report at the construction permit stage which discusses the following considerations:

a) the environmental impact of the proposed action;
b) any adverse environmental effects which cannot be avoided should the proposal be implemented;
c) alternatives to the proposed action;
d) the relationships between local short term uses of man's environment and the maintenance and enhancement of long term productivity, and
e) any irreversible and irretrievable commitments of resources which would be involved in the proposed action should it be implemented.

Of special concern here,[5]

> The Environmental Report . . . shall include a cost-benefit analysis which considers and balances the environmental effects of the facility and the alternatives available for reducing or avoiding adverse environmental effects, as well as the environmental, economic, technical and other benefits of the facility. The cost-benefit analysis shall, to the fullest extent practicable, quantify the various factors considered. To the extent that such factors cannot be quantified, this shall be discussed in qualitative terms. The Environmental Report should contain sufficient data to aid the Commission in its development of an independent cost-benefit analysis covering the factors specified in this paragraph.

The Atomic Energy Commission's Guide presents a detailed thirteen-chapter outline of the form and content expected in the applicants' environmental reports (only eleven of those chapters are substantive, however). While the cost-benefit analysis that concerns us here is strictly limited to the four final substantive chapters, the seven initial chapters are of interest because it is here that detailed information about the nature of the generating facility, its site, associated transmission facilities, and possible effects on many aspects of the human environment is presented. Much of the information presented in these chapters is to be used by the company and the Commission in reaching a determination of benefits and costs associated with the proposed facility. The information is provided to allow the company and subsequently the Commission to present a two-part cost-benefit analysis:[6]

---

[5] In 4.2.2.

> In the first part, alternative site-plant combinations and plant systems are to be examined in order to determine whether the proposed facility is the cost-effective choice, considering economic, social and other environmental factors, and any institutional (government, etc.) constraints. Because the benefit associated with each alternative will be essentially the same, cost-effectiveness analysis of the alternatives is recommended.
>
> In the second part, the benefits to be created by the proposed facility should be weighed against the aggregate of environmental, economic and other costs to be incurred.

The Guide may be divided, conceptually, into two parts. Chapters 1 through 7 outline required information to be presented by the applicant regarding the relationship of the proposed nuclear facility to the kinds of environmental problems discussed in this section. Chapters 8 through 11 present the "scientific" cost-benefit analysis that is to be employed by the applicant. The purpose of these chapters is to show that even in the presence of certain adverse environmental effects, the benefits of the proposed facility exceed the costs.

The coverage of Chapters 1 through 7 appears to be extensive. Virtually all aspects of possible environmental effects are discussed. For example, the purpose of Chapter 2 is to present a detailed discussion of the nature of the site for the plant, including the relevant information concerning land and water use, geology, hydrology, regional historic, scenic, cultural and natural landmarks, meteorology, etc. Chapter 5 is designed to present the information regarding the environmental effects of plant operation. The discussion must include both short-term and long-term effects. Included is information on the heat dissipation system and on the effects of radiation on people and other forms of life.

Chapter 7 outlines the format for discussion of the environmental effects of accidents. The AEC Guide presents a format dividing accidents into nine classes, from the most trivial to the most severe. Within each class the applicant is asked to provide information regarding accident probabilities and expected environmental impact in the context of one or more hypothesized accidents within each class suggested by the Guide. Accidents of class 1 through 8 variety are those for which safety systems have been designed for containment at some approved level (presumably established AEC safety requirements) of radiation discharge and reliability. Class 9 type accidents, ". . . involve sequences of postulated successive failures more severe than those postulated for establishing the design basis for protective systems and engineered safety features and for site evaluation purposes. Their consequences could be severe. However, the probability of their occurrence is so small that their environmental risk is extremely low." Although the Guide does not ask applicants to discuss type 9 accidents, separate hearings have been instituted to consider the associated problems.

---

[6] In 4.2.3.

Chapters 8 through 11 of the Guide present the methodology for turning the qualitative and quantitative information presented in Chapters 1 through 7 into a scientific accounting of social costs and social benefits. The necessary linkage between the first seven chapters and the cost-benefit analysis is only weakly alluded to by the Guide. Since it is the methodology of the latter chapters which concerns me most, I will outline it in somewhat more detail.

The Guide recognizes that there are limitations on the ability to evaluate the costs and benefits associated with a nuclear facility. It may be especially difficult to quantify many of the benefits and costs. Nevertheless the Guide asks for quantification wherever possible, and in Chapter 8 the methodology for calculating benefits is presented.

The Guide states that the primary benefits of the proposed facility are inherent in the value of the electricity produced. The measure of benefits indicated by the Guide is the expected average annual revenues (in dollars) of the electricity generated by the plant over the life of the plant. The year-by-year forecasts of such revenues are to be discounted to present value "using a nominal discount rate, reflecting the average cost of capital." Presumably this calculation will be accomplished by taking the kilowatt capacity of the plant, multiplying it by its expected load factor, and then multiplying the result by the number of hours in a year to get kilowatt-hour production per year. This figure would then be multiplied by the average revenue per kilowatt hour expected over the lifetime of the facility to get the annual revenue per year. These revenue figures would then be discounted back to the present to give a value for the present value of primary benefits from the plant.

Additional primary benefits to be incorporated in a similar fashion are revenues from the sale of steam, the use of waste heat for desalinization, etc.

The Guide then suggests several other social and economic benefits that may be included (in an unspecified way) in the benefit calculation. Among them are:

(1) tax revenues received by state, local and federal governments,
(2) new jobs created,
(3) increases in regional product,
(4) improved local facilities (e.g., roads), and
(5) increased knowledge of the environment, etc.

The meaningfulness of these benefit calculations will be discussed in the next section.

Chapter 8 also discusses the calculation of the costs of the proposed facility. Costs are divided into two categories—*internal* costs and *external* costs:[7]

> The primary *internal* costs are (1) the capital costs of land acquisition and improvement; (2) the capital costs of facility construction; (3) the

---

[7] In 4.2.34.

> incremental capital costs of transmission and distribution facilities; (4) fuel costs including spent fuel disposition; (5) other operating and maintenance costs including license fees and taxes; (6) plant decommissioning costs, and (7) research and development costs associated with potential future improvements of the facility and its operation and maintenance. As in the case of benefits, the applicant should discount these costs to present worth.

The applicant is then to evaluate (somehow) the long-term and short-term *external* (economic and social) costs associated with the proposed plant, presumably in the context of the environmental impact discussions presented in Chapters 4, 5, and 7. The Guide does not suggest how these things should be quantified, but merely asks that any quantification procedure used be specified and defended. I will have more to say about this chapter in the next section.

In Chapter 9, the applicant is asked to discuss alternative energy sources, including alternative sites, alternative sources of generated power (fossil fuel), and the possibility of meeting demand with no additional capacity at all. Essentially the applicant is being asked to defend his plant choice, in detail, against a set of possible alternatives. Exactly how many alternatives must be discussed is not specified, but ". . . the applicant should present its site-plant selection process as the consequence of an analysis of alternatives whose environmental costs and benefits were evaluated and compared and then weighed against those of the proposed facility." The goal of this chapter is essentially a cost-effectiveness study, since the bulk of primary benefits (revenues) are presumed to be approximately the same. So the applicant is required to list all of the internal and external costs associated with the proposed facility and its alternatives and to justify the choice that has been made. As we shall see, this chapter and the next are really the key chapters, if the methodology is to make any sense at all.

Chapter 10 extends the analysis of alternatives to the more micro level of alternative plant designs including cooling systems, waste discharge systems, and transmission facilities. The applicant is specifically requested to monetize the costs of construction and operation of these design alternatives. The environmental impact of all alternatives is to be fully documented and presumably quantified wherever possible.

Finally, in Chapter 11, the applicant is requested to put all of the information obtained so far together in the final cost-benefit calculation. Presumably, the analysis in Chapters 9 and 10 has "proven" that the choice made is cost-effective. The information from Chapters 4, 5, and 7 has been used to estimate internal and external costs in Chapter 8, so that the company need only add up the costs and benefits as calculated in Chapter 8 to see whether the benefits outweigh the costs. Presumably all reports submitted by serious applicants will show positive net benefits.

While the suggested methodology appears to seek precision in the accounting of benefits and costs, the question remains as to whether the accounting procedure really speaks to the environmental problems in a meaningful way.

4. Critique of the AEC cost-benefit methodology

■ When this Guide is read for the first time, one's gut reaction is that of horror. If all major construction projects in the economy had to go through this kind of administrative review, would anything ever get built? What are the bureaucrats going to do with all of this information? Is the supposed cost-benefit analysis really anything more than a tableau that lays out comparisons between apples and oranges? Does this type of approach really make private construction decisions respond to environmental costs or merely cause incredible administrative delay? More careful reflection leads me to believe that some of these initial fears may be well-founded. The proposed cost-benefit analysis has severe methodological as well as practical flaws. More importantly, this whole approach to environmental protection may generate more paper work than protection. Finally, nuclear power may present evaluation problems for which there are no well-established or acceptable evaluation criteria.

Of primary importance is the question of whether cost-benefit analysis even makes sense in this situation. Traditionally, cost-benefit analysis evolved to deal with situations in which various types of market failures made the provision of particular types of services, *combined* with efficient pricing, difficult or impossible in private markets.

Depuit's bridge, for example, is a lumpy investment situation in which efficiency dictated very low prices (in the absence of crowding) equal to short-run marginal cost. But efficient pricing (assuming perfect price discrimination is impossible) would mean negative profits, and no provision of service by a private market. Cost-benefit analysis would then be used to help the government to decide whether a subsidy (or direct provision of the service) combined with marginal cost pricing was justified. Similarly, in various water resource, urban renewal, and research and development situations, many benefits may not be captured by a private competitive market. Again, theoretically, the government might use cost-benefit analysis to determine whether positive net social benefits indicated the desirability of government subsidies for the projects in question, or even direct government investment. The general tools of cost-benefit analysis have been developed to deal with situations in which private competitive markets yield *suboptimal* levels of activities in terms of potential social benefits.

The environmental evaluation context under discussion here is almost the inverse of the traditional cases. Here we have private (regulated monopoly) firms that want to provide a particular service (electricity) by using a particular production technique. However, society (through NEPA) is fearful that *too much* of the service will be provided and that socially inefficient production techniques will be employed. This fear arises because it is believed that there are external costs associated with the production of the service in question that lead both to too much of the service's being produced (presumably because output prices are too low) and the wrong production techniques' (because various inputs into the production process are underpriced, leading to cost-minimizing production decisions that are not socially optimal) being chosen. This problem has of course been analyzed in the context of externalities.

The Atomic Energy Commission has chosen to deal with this latter problem through environmental reports in the form of cost-benefit analyses. On the surface then, Chapter 8 of the Guide appears to be the key to the AEC decisionmaking process. The applicant is asked to compare the present discounted value of future revenues (plus a number of other benefits of questionable theoretical relevance)[8] with the present discounted value of *incurred* costs plus *external* costs. The incurred costs are presumably those of the most *socially* efficient production process as determined in Chapters 9 and 10. (Exactly why the cost-effectiveness analysis appears after the cost-benefit calculation is a mystery, since without it the cost-benefit analysis is entirely meaningless.) But this in itself is quite odd, since regulators are presumably setting utility prices so that the present discounted value of revenues is equal to the present discounted value of incurred costs. If rate regulators are doing their job properly, the AEC formula should always yield costs in excess of benefits if there is even $1 of "external" costs!

The problem arises because the key calculations for efficient production decisions do not appear in the cost-benefit calculations of Chapter 8, but rather in the cost-effectiveness analysis of Chapters 9 and 10. Presumably, the company has displayed all of the alternative techniques for the production of electricity, including the *incurred* costs and *residual* externalities associated with each. In making its choice, the company must have some implicit valuation scheme it put on the externalities *or* a recognition of effluent standards specified by law (or understanding with the AEC) *or* a combination of both. Similarly, in giving its approval to the production technique chosen, the AEC must have some implicit set of values it puts on the external costs or some set of effluent standards it is using or a combination of both. Otherwise, comparisons of alternative production techniques yielding different combinations of incurred and external costs would be impossible.

A socially efficient production technique normally implies a certain level of *incurred* costs per unit of output *plus* residual environmental costs.[9] The combination of incurred costs plus residual environmental costs is efficient in the sense that it represents the least-cost method of production available, given the social (shadow) prices on all relevant inputs and outputs and the recognition of established standards where social shadow prices have not been calculated (or a combination of both). Let us represent the relationship between output and long-run marginal social costs (calculated in this way) as MSC_1 in Figure 1. Let us represent the long-run marginal costs *incurred* by MCI_1 (excluding the cost of residual externalities). A demand function is also indicated in Figure 1 to represent the *additional* demand that the utility is seeking to meet by building the proposed plant. (For simplicity I use a static model, but we may think of the cost curves as being present discounted costs and the demand functions as being present discounted revenues.) Recall that MSC_1 and MCI_1 have been arrived at through the cost-effectiveness analysis of Chapters 9 and 10.

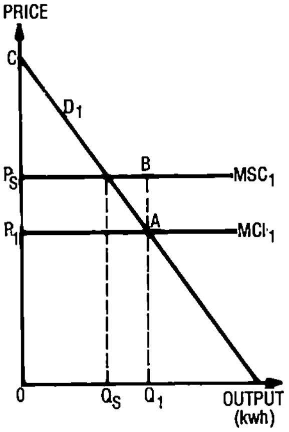
---

[8] For example, it does not appear to be correct to include tax revenues as additional benefits, since they are normally included when the regulatory commission sets output prices. Taxes will then be counted twice. Similarly, increases in regional product and incomes are presumably what is causing the increased demand for electricity. The value of increased regional product attributable to electricity will already be reflected in the expected revenue calculation.

[9] It is conceivable that the efficient solution will involve no environmental costs, but this is unlikely. The important consideration is whether *all costs* have been properly taken into consideration by the private market when production techniques are chosen and final output prices arrived at, i.e. *external* costs have been *internalized* to the decisions of private firms.

FIGURE 1

PRICES BASED ON INCURRED COSTS VERSUS PRICES BASED ON UNCHARGED "SOCIAL" COST



Given MSC_1 and MCI_1, we then go to Chapter 8 to compare revenues with costs. However, MSC_1, the true social marginal cost curve, is only a phantom. It is a construct of the cost-benefit analysis, because the firm has not actually been *charged* with the value of residual environmental costs. Ideally, we expect regulators to set price equal to the marginal costs actually incurred ($P_1 = MCI_1$) and the firms will expect output Q_1 to be demanded. Total revenues will be the rectangle P_1AQ_10. *As a result, if there are any residual environmental costs, the AEC methodology, if rigorously applied, should lead to the rejection of the project.*

If we were central planners, we would specify the socially efficient production technique arrived at in Chapters 9 and 10, but would set a price P_S yielding a demand of Q_S. Projects would be accepted, but on a smaller scale. Construction would only be rejected entirely if $MSC_1 > C$. If we assume that P_1 represents the prevailing price for electricity based on the costs of construction and operation being incurred for existing demand, Q_1 additional units of output will be demanded, but the AEC methodology will only accept construction if $MSC_1 = MCI_1$. There will inevitably be shortages resulting from this approach. At best, the cost-effectiveness analysis of Chapters 9 and 10 will indicate nuclear power as the socially efficient production technique and Chapter 8 will reject it, because the calculated benefits will be equal only to the incurred costs. In addition, if regulators base prices on incurred costs only, there will be excess demand that will no doubt be met by building fossil-fuel plants (not requiring AEC approval). But the cost-effectiveness analysis had supposedly indicated that the fossil-fuel production technique as not socially efficient. Catch-22!

Even more bizarre are the results which emerge if public utility regulators do a bad job setting prices. It has been argued by some that public utility commissions do not really force prices all the way down to average cost (equal to marginal cost in this example), but that the companies are allowed to earn excess profits: $P_1 > MCI_1$. Strangely enough, by allowing the company to earn excess profits, the utility commissions may (unwittingly) allow the AEC methodology to accept projects which have residual environmental costs (where $P_1 \geqslant MSC_1$). If the utility commission fortuitously sets $P_1 = MSC_1$ instead of $P_1 = MCI_1$, and if the cost-effectiveness study has chosen the proper production technique, the resulting equilibrium is socially efficient and the AEC methodology will (properly) accept the proposed project. We have then a possibility of countervailing bureaucratic mistakes, yielding an efficient solution. It is doubtful that we want to count on state utility commissions to do a bad enough job to bail the AEC methodology out. Some may also be concerned about the distributional implications of the fortuitous socially efficient equilibrium that results.

The primary problem is that cost-benefit analysis is probably not the proper tool for effecting environmental protection in this situation (or probably in any others). We may return to the proposed cost-effectiveness analysis of Chapters 9 and 10 of the Guide to see exactly why.

In Chapters 9 and 10, the applicant is asked to lay out the alternative sites, plant types, plant designs, etc., then to monetize the costs and residual externalities associated with each. We would normally expect a private firm to examine all input prices (plus safety and effluent standards) and to pick the production technique that yields the minimum *incurred* costs (subject to any effluent and safety standards). This is probably exactly what the utilities have done in choosing their proposed production technique. The *raison d'être* for this entire Guide—Chapters 9 and 10 in particular—is the feeling that all scarce inputs have not been assigned their proper social prices or (in the absence of shadow prices on all inputs) the proper effluent and safety standards have not been instituted. The AEC presumably examines this list of alternatives and applies its own valuations and/or safety and discharge standards to see whether the socially cost-minimizing technique has been chosen. If, under its own valuation scheme, the company had not chosen the cost-minimizing technique, then changes would be asked for, and the acceptable technique would then be subject to the cost-benefit analysis of Chapter 8, with the same problems already indicated.

But this seems to be an extremely inefficient way of going about things. If the AEC has specific safety and effluent standards that it is using as evaluation criteria, why does it not simply specify them and tell the company to design a system that meets the standards? If it has specific shadow prices in mind for other external effects, why does it not seek to implement a system that imposes these shadow prices on inputs used in the production process? Faced with a set of standards and shadow prices, the company, which no doubt knows much more about the production processes than the AEC, would then simply pick the cost-minimizing technique, given the social input prices and the imposed standards. The firm would now be forced to face the proper social resource tradeoffs, and the costs actually *incurred* would be *social* costs. Final output prices would equal marginal social cost and the firm would only seek to construct as much production capacity as it could profitably sell, given prevailing demand characteristics.

In a sense then, the cost-benefit analysis is an unnecessary tool. If the AEC really has the information to evaluate social costs properly, then it would do better to apply this information directly to incorporate externalities into the private decisionmaking process. The role of the AEC would then be one of setting standards, evaluating external costs, monitoring the behavior of the nuclear plants, enforcing safety and effluent standards, and placing shadow prices on other externalities.

My suspicion is that, in reality, the AEC is doing little more than this basic monitoring function. The filings under the environmental guide provide detailed qualitative and quantitative information about nuclear plants, site location, design alternatives, and externalities, which may in the long run provide the capability for effecting a meaningful set of standards and shadow prices. In the meantime, the cost-benefit analysis itself may be little more than an attempt to respond to Skelly Wright's decision and to create enough administrative delay to postpone massive introductions of nuclear generating facilities. To the extent that the cost-benefit analysis is merely a charade we may be better off, for if the AEC methodology were rigorously applied, it would lead to some ridiculous results.

5. Low probabilities and large losses

■ The massive amounts of pseudoscientific cost-benefit jargon contained in the Guide do obscure two important policy problems, however. If we are going to attempt to deal effectively with externalities in the context of private markets, we must deal with them directly, either through the imposition of appropriate shadow prices or a set of rules and standards[10] that will effectively internalize them to the private decisionmaking processes. It would be most unfortunate, if the "cost-benefit" approach presented by the AEC Guide were to become the generally accepted method of dealing with environmental costs through NEPA. The cost-benefit approach requires at least as much information as would direct efforts aimed at the internalization of the externalities, and once the internalization is done, the cost-benefit analysis is no longer necessary. It is the direct approach rather than the cost-benefit approach that we should aim for.

While these observations are of general validity in the application of economic theory to environmental problems, nuclear power may be a difficult case to deal with. Many of the social costs associated with nuclear power involve situations in which one or more of the following is the case:

(1) Insufficient information about the long-term effects on environment of even relatively low levels of radioactivity (what Frank Knight would call uncertainty).
(2) Unsatisfactory information about the reliability of existing safety systems. Many of the reliability estimates are based on computer models whose component parts have empirical validity, but which, taken as a whole, have been subject to little empirical verification.
(3) Very small probabilities of very large and long-term losses.

These three phenomena taken together make it very difficult to make reliable estimates of the costs of massive reliance on nuclear power for the generation of electricity. (1) and (3) make it difficult to put reactors through the rigorous kinds of shakedown and tolerance exercises that more traditional pieces of equipment go through, thus making it difficult to improve our knowledge of (2). In addition, established methods for evaluating uncertain events in the context of low probabilities of very large losses may be inadequate when the human stakes are immense.

---

[10] See Weitzman [3].

Unless we can come up with reliable information about the effects of radioactivity on human beings and other living things and have more empirical information about reactor reliability and safety, not only is the AEC type of cost-benefit analysis questionable, but the direct attempt to internalize the social costs through shadow prices and/or standards may be impossible. Finally, even given solutions to problems (1) and (2), some kind of decision will have to be made by society concerning how it wishes to value the high-loss, low-probability accidents.

In conclusion, it appears that the AEC Guide presents a cost-benefit structure that is largely a facade. These filings will help the AEC to gain more information about the alternative production processes, their costs, and their environmental effects. In the long run, this information will be valuable for dealing with environmental problems directly through the use of social shadow prices or specific standards. In the short run the result will probably be substantial delays in the construction and operation of nuclear plants, forcing many utilities to build additional fossil-fuel capacity to satisfy the demand for electricity. While these delays buy time to obtain further information about the environmental impacts of nuclear power, this time is not costless. Society pays for it by using alternative fossil-fuel technologies (having their own environmental problems) which may, in retrospect, turn out to involve higher total social costs than does nuclear power.

Recent statements by Nixon Administration officials indicate that they do not find the current status of this tradeoff to be satisfactory, given the new realities of fossil-fuel supplies. As a result, we can look forward to efforts aimed at speeding up the approval process. Hopefully, this program will be accompanied by increased efforts to ascertain the environmental effects of atomic energy and to incorporate effectively the consideration of the associated costs in the decision process.

References

1. *Calvert Cliffs' Coordinating Committee, Inc. v. Atomic Energy Commission,* 449 F. 2d 1109, D.C.C.A., 1971.
2. U. S. Atomic Energy Commission. *Preparation of Environmental Reports for Nuclear Power Plants.* Washington, D. C., 1973.
3. Weitzman, M. "Prices vs. Quantities." M.I T. Working Paper No. 106, April 1973.

THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE

EDITOR
Paul W. MacAvoy, *Massachusetts Institute of Technology*

ASSOCIATE EDITORS:
Dale W. Jorgenson, *Harvard University*
Harvey J. McMains, *American Telephone and Telegraph Company*
Oliver E. Williamson, *University of Pennsylvania*

ASSISTANT TO THE EDITOR:
Leigh E. Tripoli, *American Telephone and Telegraph Company*

AUTUMN '74
VOL. 5. NO 2.

TABLE OF CONTENTS

ARTICLES

335 Theories of economic regulation
R. A. POSNER

359 Free riders and collective action: an appendix to theories of economic regulation
G. J. STIGLER

366 Accident law and resource allocation
P. A. DIAMOND

406 The choice between capitalizing and expensing under rate regulation
P. B. LINHART, J. L. LEBOWITZ and F. W. SINDEN

420 Alternative investment models for firms in the electric utilities industry
S. C. PECK

SYMPOSIUM ON ANALYTICAL MODELS OF ENERGY POLICY

✓ 461 U.S. energy policy and economic growth, 1975-2000
E. A. HUDSON and D. W. JORGENSON

515 The effects of higher prices on electricity consumption
J. M. GRIFFIN

✓ 540 An economic model of the world oil market
M. KENNEDY

✓ 578 Energy, environment, and economic growth
A. P. CARTER

SHORT ARTICLES, COMMENTS AND REVIEWS

595 Railroad diversification: where lies the public interest?
G. C. EADS

614 The losses on savings deposits from interest rate regulation
D. H. PYLE

623 The implications of regulation for induced technical change
V. K. SMITH

633 The regulatory implications of three alternative econometric supply models of natural gas
R. S. PINDYCK

646 Determinants of research and development activity by electric utilities
R. P. WILDER and S. R. STANSELL

651 Dynamic aspects of public goods: a further analysis of the telephone system
B. von RABENAU and K. STAHL

670 Insurance as a market in contingent claims: structure and performance
J. M. MARSHALL

683 The economics of product safety and imperfect information
V. P. GOLDBERG

689 The economics of product safety: a rejoinder
W. Y. OI

696 On "duration" and the optimal maturity structure of the balance sheet
M. A. GROVE

710 The behavior of a firm subject to stochastic regulatory review: comment
W. T. ZIEMBA

713 The behavior of a firm subject to stochastic regulatory review: correction
A. K. KLEVORICK

715 Lieb's *Freight Transportation: A Study of Federal Intermodal Ownership Policy*
H. A. HOROWITZ

Theories of economic regulation

Richard A. Posner

Professor of Law
University of Chicago
and Senior Research Associate
National Bureau of Economic Research

Several theories have been advanced to explain the observed pattern of government regulation of the economy. These include the "public interest" theory and several versions, proposed either by political scientists or by economists, of the "interest group" or "capture" theory. This article analyzes those theories. It argues that the public interest theory and the political scientists' versions of the interest group theory are unacceptable in their present form. The economists' version of the interest group theory is discussed at greatest length; its theoretical and empirical foundations are reviewed, and the conclusion is reached that, while promising, the theory requires both more analytical development and new sorts of empirical investigation before it can be accepted as an adequate positive theory of regulation.

1. Introduction

■ A major challenge to social theory is to explain the pattern of government intervention in the market—what we may call "economic regulation." Properly defined, the term refers to taxes and subsidies of all sorts as well as to explicit legislative and administrative controls over rates, entry, and other facets of economic activity. Two main theories of economic regulation have been proposed. One is the "public interest" theory, bequeathed by a previous generation of economists to the present generation of lawyers.[1] This theory holds that regulation is supplied in response to the demand of the public for the correction of inefficient or inequitable market practices. It has a number of deficiencies that we shall discuss. The second theory is the "capture" theory—a poor term but one that will do for now. Espoused by an odd mixture of welfare state liberals, muckrakers, Marxists, and free-market economists, this theory holds that regulation is supplied in response

---

Richard A. Posner received the A.B. from Yale University in 1959 and the LL.B. from Harvard University in 1962. His current research centers on the use of economic analysis to explain the behavior of the legal system.

The author wishes to thank Gary S. Becker, Ronald H. Coase, Paul W. MacAvoy, B. Peter Pashigian, and George J. Stigler for their helpful comments on an earlier draft.

This paper was originally given at a conference on regulation sponsored by the National Bureau of Economic Research, and is part of a series of studies in the area of law and economics being carried out by the Bureau under a grant from the National Science Foundation. It is not an official Bureau publication, since it has not yet undergone the full critical review accorded National Bureau publications, including approval by the Bureau's board of directors.

For an appendix to this article, see George Stigler's "Free Riders and Collective Action," pp. 359–365 in this issue.

[1] The theory is more often assumed than articulated. Some representative works embodying it are Bonbright [4], Davis [9], and Friendly [13].

to the demands of interest groups struggling among themselves to maximize the incomes of their members. There are crucial differences among the capture theorists. I shall argue that the economists' version is the most promising, but shall also point out the significant weaknesses in both the theory and the empirical research that is alleged to support the theory.

2. The public interest theory of regulation

■ **The original theory.** Two assumptions seem to have typified thought about economic policy (not all of it by economists) in the period roughly from the enactment of the first Interstate Commerce Act in 1887 to the founding of the *Journal of Law and Economics* in 1958. One assumption was that economic markets are extremely fragile and apt to operate very inefficiently (or inequitably) if left alone; the other was that government regulation is virtually costless. With these assumptions, it was very easy to argue that the principal government interventions in the economy—trade union protection, public utility and common carrier regulation, public power and reclamation programs, farm subsidies, occupational licensure, the minimum wage, even tariffs—were simply responses of government to public demands for the rectification of palpable and remediable inefficiencies and inequities in the operation of the free market. Behind each scheme of regulation could be discerned a market imperfection, the existence of which supplied a complete justification for some regulation assumed to operate effectively and without cost.

Were this theory of regulation correct, we would find regulation imposed mainly in highly concentrated industries (where the danger of monopoly is greatest) and in industries that generate substantial external costs or benefits. We do not. Some fifteen years of theoretical and empirical research, conducted mainly by economists, have demonstrated that regulation is not positively correlated with the presence of external economies or diseconomies or with monopolistic market structure. Few, if any, responsible students of the airline industry, for example, believe that there is some intrinsic peculiarity about the market for air transportation that requires prices and entry to be fixed by the government. The same may be said for trucking, taxi service, stock brokerage, ocean shipping, and many other heavily regulated industries. Even the danger of "market failure" in such traditionally unquestioned areas of regulation as health care, the legal profession, and the safety of drugs and other products is increasingly discounted. The conception of government as a costless and dependably effective instrument for altering market behavior has also gone by the boards.[2] Theoretical revision has both stimulated and been reinforced by a growing body of case studies demonstrating that particular schemes of government regulation—whether of taxicabs, or producers of natural gas, or truckers, or airlines, or stock brokers, or new drugs, or electricity rates, or broadcasting—cannot be explained on the ground that they increase the wealth or, by any widely accepted standard of equity or fairness, the justice of the society.[3]

---

[2] Some examples of the emerging theory of "government failure" are Hirshleifer, DeHaven, and Milliman [16], pp. 74–82, and Posner [39] and [35].

☐ **A reformulation.** The empirical evidence is sometimes challenged on the ground that the disappointing performance of the regulatory process is the result not of any unsoundness in the basic goals or nature of the process but of particular weaknesses in personnel or procedures that can and will be remedied (at low cost) as the society gains experience in the mechanics of public administration.[4] Thus reformulated, the public interest theory of regulation holds that regulatory agencies are created for bona fide public purposes, but are then mismanaged, with the result that those purposes are not always achieved.

This reformulation is unsatisfactory on two grounds. First, it fails to account for a good deal of evidence that the socially undesirable results of regulation are frequently desired by groups influential in the enactment of the legislation setting up the regulatory scheme. The railroads supported the enactment of the first Interstate Commerce Act, which was designed to prevent railroads from practicing price discrimination, because discrimination was undermining the railroads' cartels. American Telephone and Telegraph pressed for state regulation of telephone service because it wanted to end competition among telephone companies. Truckers and airlines supported the extension of common carrier regulation to their industries because they considered unregulated competition "excessive." Sometimes the regulatory statute itself reveals an unmistakable purpose of altering the operation of markets in directions inexplicable on public interest grounds, as in the reference in the ICC's statutory mandate to the desirability of maintaining "balance" among competing modes of transportation.[5] None of this evidence is decisive against the public interest theory—in each case other groups besides the industry directly regulated supported the legislation. Whether the other groups were also interest groups is discussed later on.

Second, the evidence that has been offered to show mismanagement by the regulatory agency is surprisingly weak. Much of it is consistent with the rival theory (which is considered more closely in Section 3) that the typical regulatory agency operates with reasonable efficiency to attain deliberately inefficient or inequitable goals set by the legislature that created it. The proclivity of some agencies for concentrating their resources heavily on cases of small individual consequence—a proclivity often thought to be convincing evidence of mismanagement—is in fact consistent with an efficient allocation of resources within the agency.[6] The frequent criticisms of agencies for relying on case-by-case adjudication to make policy, rather than engaging in elaborate planning exercises, are extremely superficial since they ignore, first, the intrinsic difficulty of forecasting the future and, second, the disastrous consequences for agencies, notably the Federal Communications Commission, that have engaged in such planning.[7] The common argument that the employees of regulatory agencies must be less able than their counterparts in the private sector, since they are paid lower salaries,[8] ignores the fact that service with an agency frequently increases the later earning capacity of the employee in the private sector. The agency makes a contribution to the employee's human capital. This contribution, when added to his salary, may equal the value of the salary (plus contributions of human capital) that he would have received in the private sector.[9] In sum, one is left puzzled as to why such failures of regulation in the public interest as one observes should be ascribed so confidently to bureaucratic ineptitude.

---

[3] See, for example, Baxter [1], Cabinet Task Force on Oil Import Control [6], Coase [7], Hilton [15], Jordan [19], Kitch, Isaacson, and Kasper [21], MacAvoy [25], Peltzman [30, 31], and, for a general summary of the literature, Jordan [20].

[4] An interesting example of this point of view is provided by Herring in [14], which argues that the challenge to public administration is to develop techniques of overcoming the interest group pressures that threaten to deflect legislative programs from serving the public interest. He may underestimate the difficulties of doing this because of a certain economic naivete that leads him to suggest that opposing group interests can often be harmonized in a way that vindicates the public interest. Ordinarily harmony is achieved at the expense of the public interest. See text at notes 38–41, *infra.*

[5] National Transportation Policy, 49 U.S.C. preceding § 1.

Third, no persuasive theory has yet been proposed as to why agencies should be expected to be less efficient than other organizations. The motivation of the agency employee to work diligently and honestly is similar to that of the employee of a business firm. Both want to obtain advancement (not necessarily within the employing firm or agency) and to avoid being fired, demoted, or humiliated. To some extent, these motivations are independent of the incentive of the agency's head to enforce standards of diligence and honesty against the employees. Many employees will want to demonstrate the possession of excellent qualities in order to improve their prospects for superior private employment anyway. In any event, the agency head's incentive is clear. He derives few benefits from the slackness of his staff—not even the famous "quiet life." His life would not be so quiet, for many employees would be restless and dissatisfied, knowing that their opportunities for private employment were being impaired by the agency's reputation for laxity and sloth.

Furthermore, the agency's head is answerable both to the legislative and (if he desires promotion or reappointment) to the executive branches. Legislative oversight of agencies is too little emphasized. Unlike business firms, government agencies must go to *their* capital markets—the legislative appropriations committees —every year. There is competition among agencies for the largest possible slice of the appropriations pie, and the agency that has a reputation for economy and hard work enjoys an advantage in the competition, for only in the exceptional case will it be to the legislators' advantage that the agency's personnel be lining *their* pockets (whether with pecuniary income or with nonpecuniary income such as leisure).[10]

---

[6] Posner [33].

[7] See, for example, Posner [32] and Comanor and Mitchell [8].

[8] The question is never asked whether their jobs might be less risky or require less skill than the private-sector jobs with which they are compared.

[9] See Posner [34] and Eckert [11].

One objection to the foregoing argument is that the agency differs from the private firm in not competing in any product market. But that is to say only that the agency is like a private monopolist, and there is no convincing theoretical or empirical support for the proposition that the internal management of monopolistic firms is any laxer than that of competitive firms. Another objection is that the agency has little incentive to minimize costs because, unlike a business firm, it cannot keep the profits generated by its cost savings. Yet most employees of business firms do not share in the profits of the enterprise, and they are somehow motivated to work efficiently. Moreover, I have suggested several ways in which agency employees, from the head of the agency down, do "profit" from efficient management, and lose if the agency is managed inefficiently.

☐ **A further reformulation of the public interest theory.** The idea that regulation is an honest but frequently an unsuccessful attempt to promote the public interest becomes somewhat more plausible if we introduce two factors often ignored. The first is the intractable character of many of the tasks that have been assigned to the regulatory agencies. The clearest example is the regulation of price levels under public utility and common carrier statutes. These statutes require the agencies to determine the costs of the regulated firms and to hold their prices to those costs, and there are good grounds for believing that the necessary instruments of measurement and control simply do not exist.[11] The agencies are asked to do the impossible and it is not surprising that they fail, and in attempting to succeed distort the efficient functioning of the regulated markets. But this does not explain why legislatures assign such tasks to agencies.

The second factor is the cost of effective legislative supervision of the agencies' performance. In a recent article on legal rulemaking, Isaac Ehrlich and I point out that legislative bodies are a type of firm in which the costs of production are extremely high and, moreover, rise very sharply with increases in output.[12] The reason is that legislative "production" is a process of negotiation among a large group, the legislators, and the analysis of transaction costs in other contexts suggests that bargaining among a number of individuals is a costly process (and explains why legislatures require only a majority and not a unanimous vote in the conduct of their business). Because costs of bargaining rise rapidly with the number of bargainers, a legislature cannot respond efficiently to a growth in workload by increasing the number of its members. Hence, as the business of a legislature rises, it can be expected to delegate more and more of its work to agencies, and to exercise progressively less control over those agencies. This theory has various testable implications. It suggests, for example, a "life cycle" theory of administrative regulation. The agency is created at a time when the legislature has a strong interest in the problem to be dealt with by the agency. But as time passes, and other problems come before the legislature, the legislature finds itself unable at reasonable cost to continue to devote time to properly monitoring the agencies created previously. The theory also implies that administrative failure will become, on average, a more serious problem over time, with the growth of the size and complexity of the economy. As we shall see shortly, however, the inquiries suggested by these hypotheses might not discriminate adequately between the version of the public interest theory suggested here and some versions of the capture theory of regulation.

---

[10] One could argue that the legislator may not have much incentive to ride herd on the agency: he will not get paid more and his popularity with the voters will be increased only marginally. But this ignores the fact that the actual audit will be conducted by an employee of the legislative body, who will have the same incentive to conduct a searching audit as any privately employed auditor.

[11] See, for example, Posner [35].

[12] Ehrlich and Posner [12].

☐ **Behavioral assumptions of the public interest theory.** A serious problem with any version of the public interest theory is that the theory contains no linkage or mechanism by which a perception of the public interest is translated into legislative action. In the theory of markets, it is explained how the efforts of individuals to promote their self-interest through transacting bring about an efficient allocation of resources. There is no comparable articulation of how a public perception as to what legislative policies or arrangements would maximize public welfare is translated into legislative action. It is not enough to say that a voter will vote for the candidate who promises to carry out the policies that the voter perceives to be in the public interest; other policies might benefit the particular voter more. Policies that benefitted 51 percent of the voters might impose much greater costs on the other 49 percent, in which event the majority would be confronted with a conflict between principle and interest—and no body of theory or of evidence suggests that they would be likely to vote the former.

There are two possible ways around this problem. One, suggested by Ronald Coase, emphasizes the moral differences between private and political action. The assumption that market behavior is normally motivated by fairly narrow considerations of self-interest is plausible, because most market decisions are social goods rather than bads. To be sure, a decision to sell a new product may harm a competitor or a locality or a group of workers or of customers, but the decision makers can be reasonably confident that these harms are more than offset by the gains to others. Where, however, an individual votes for policies designed to exploit his fellows, he can hardly avoid confronting the moral implications of his action and the moral code may constrain him from voting in that manner.

A second approach is to observe the potentiality for collusion among politicians. There are only two important political parties in this country, and there are barriers not only to the formation of additional parties but to the takeover of either of the major parties by disgruntled members or outsiders. Thus, there would appear to be opportunities for the politicians who dominate the parties to agree to impose some of their own policy preferences on the electorate. They could also use their monopoly power to obtain pecuniary income—and doubtless do—but I am assuming that they take at least some of their monopoly profits in the form of satisfaction from imposing on the public their conception of the public interest (which might differ from the conception held by the electorate and from the desires of any particular interest group). If this analysis is accepted, it becomes plausible to suppose that some policies are adopted because they conform to the public interest—as conceived by the politicians.

3. Some versions of the capture theory

■ **The Marxists and the muckrakers.** The theory that economic regulation is not about the public interest at all, but is a process by which interest groups seek to promote their (private) interests, takes several distinct forms. One, which is put forward by Marxists and by Ralph Nader-type muckrakers, can be crudely summarized in the following syllogism. Big business—the capitalists—control the institutions of our society. Among those institutions is regulation. The capitalists must therefore control regulation. The syllogism is false. A great deal of economic regulation serves the interests of small-business—or nonbusiness—groups, including dairy farmers, pharmacists, barbers, truckers, and, in particular, union labor. Such forms of regulation are totally unexplained (and usually either ignored or applauded) in this version of the interest-group or "capture" theory.

☐ **The political scientists' formulations.** A more interesting version of the "capture" theory derives from political science, and in particular from Bentley and Truman and their followers, who emphasize the importance of interest groups in the formation of public policy.[13] The political scientists have developed some evidence of the importance of interest groups in legislative and administrative processes, but unfortunately their work is almost entirely devoid of theory. They do not tell us why some interests are effectively represented in the political process and others not, or under what conditions interest groups succeed or fail in obtaining favorable legislation.[14]

A few political scientists have proposed the rudiments, at least, of a usable theory. This theory—which the term "capture" describes particularly well—is that over time regulatory agencies come to be dominated by the industries regulated.[15] This formulation is more specific than the general interest group theory. It singles out a particular interest group—the regulated firms—as prevailing in the struggle to influence legislation, and it predicts a regular sequence, in which the original purposes of a regulatory program are later thwarted through the efforts of the interest group.

---

[13] See Bentley [2] and Truman [44]. For a specific application of the approach to regulatory agencies, see Truman [44], pp. 416–421. For the position of Herring on the role of interest groups in regulation see note 4, *supra*. It is interesting to note that "interest group" is not a pejorative term for most of the political scientists, since they are either indifferent to or unaware of the fact that the economic costs of regulation procured by an interest group normally exceed the economic benefits.

[14] The vagueness of the theorizing in Truman's book [44], especially pp. 506–507 and 515, is typical.

[15] See Bernstein [3], Huntington [17], Leiserson [23], and Ziegler [46].

Unfortunately, the theory is still unsatisfactory. First, it is confusingly similar to, and in practice probably indistinguishable from, some versions of the public interest theory discussed in Section 2. Second, while I have generously called it a "theory," it is actually a hypothesis that lacks any theoretical foundation. No reason is suggested for characterizing the interaction between the regulatory agency and the regulated firm by a metaphor of conquest, and surely the regulatory process is better viewed as the outcome of implicit (sometimes explicit) bargaining between the agency and the regulated firms. No reason is suggested as to why the regulated industry should be the only interest group able to influence an agency. Customers of the regulated firm have an obvious interest in the outcome of the regulatory process—why may they not be able to "capture" the agency as effectively as the regulated firms, or more so? No reason is suggested as to why industries are able to capture only existing agencies—never to procure the creation of an agency that will promote their interests—or why an industry strong enough to capture an agency set up to tame it could not prevent the creation of the agency in the first place.

The "theory" answers none of these questions. In addition, it is contradicted by three important bodies of evidence. First, not every agency is characterized by a pristine virtue; often there is no occasion for conquest. As mentioned earlier, there is now considerable evidence that a major purpose (in fact) of the original Interstate Commerce Act was to shore up the railroads' cartels.[16] Later amendments, typically passed at the behest of the Commission itself, seem to have been less rather than more favorable to railroads (an example is the Hepburn Act which gave the ICC the power to fix maximum rates). The sequence is opposite to what the capture hypothesis predicts.

Second, the theory has no predictive or explanatory power at all when a single agency regulates separate industries having conflicting interests. The ICC is again a conspicuous example. It regulates competing modes of transportation—truckers, railroads, and barge lines—and the theory does not tell us which one the ICC can be expected to favor. This difficulty is not limited to the agency with a multiindustry "clientele." There are always competing groups within an industry. The interests of the trunk airlines are not identical to those of the regional or of the local service lines: which will the CAB decide to promote? The interests of the telephone companies, primarily AT&T, are in conflict with those of Western Union and other "record" carriers: which competing group will the Federal Communications Commission promote?

Third, the capture theory ignores a good deal of evidence that the interests promoted by regulatory agencies are frequently those of customer groups rather than those of the regulated firms themselves. Indeed, not only many examples of specific regulatory policies, but some of the structural characteristics of the regulatory process, seem best explained by reference to the influence on the regulatory process of interest groups consisting of customers of the regulated industry.[17]

---

[16] See Hilton [15], Kolko [22], and MacAvoy [24].

☐ **The economic theory of regulation.** What I shall call "the economic theory of regulation" was proposed by George Stigler in a pathbreaking article.[18] The theory seems at first glance merely a refined version of the capture theory just discussed. It discards the unexplained, and frequently untrue, assumption of pristine legislative purpose; it admits the possibility of "capture" by interest groups other than the regulated firms; and it replaces the "capture" metaphor, with its inappropriately militaristic flavor, by the more neutral terminology of supply and demand. But it insists with the political scientists that economic regulation serves the private interests of politically effective groups.

More is involved, however, than merely a recasting of the work of the political scientists. The economic theory is more precise and hard-edged—easier to confront and test with a body of data—than the political theory (which, as I pointed out, is not really a theory at all). Moreover, the economic theory is committed to the strong assumptions of economic theory generally, notably that people seek to advance their self-interest and do so rationally. A political scientist can argue that regulation is more likely to be imposed in a declining industry because adversity is a greater spur to effort than opportunity[19] (an example that assumes that regulation is normally obtained for the benefit of the regulated firms). The economist is reluctant to accept such an explanation. He does not distinguish between a profit foregone and a loss incurred—the former is a cost too, indeed the same kind of cost.[20] (I note parenthetically that the hypothesis is contradicted by a good deal of evidence.[21])

It is, of course, a weakness rather than a strength in a theory that it is so elastic as to fit any body of data with which it is likely to be confronted. The political science theory of regulation is such a theory. Exceptions to the general rule that regulatory agencies are captured by the regulated firms are explained away by facile references to the personality of the legislators, public opinion, ignorance, folk wisdom,[22] etc. The economic theory insists that regulation be explained as the outcome of the forces of demand and supply. Outcomes that cannot be so explained count as evidence against the theory.

---

[17] See Posner [40].

[18] Stigler [42]. For an attempt to marshal empirical support for the theory see Jordan [20]. Stigler, it should be noted, builds on earlier work by economists on the political system. See Buchanan and Tullock [5], Downs [10], and Olson [28].

[19] Wilson [45].

[20] Coase, however, suggests the interesting possibility that when a business is in decline, managers may find it profitable to shift their attention from improving their business operations to improving the political environment, another potential source of profits. There is a further point: the costs of becoming informed about opportunities for enhanced profits through government regulation may be greater than the costs of perceiving losses that regulation might reduce.

[21] For example, the airline industry was not declining in 1938 when the Civil Aeronautics Act was passed, nor the railroad industry in 1887 when the first Interstate Commerce Act was passed. One tends to associate regulation with declining industries primarily because so many regulatory programs were instituted during the depression of the 1930s.

4. A closer look at the economic theory of regulation

■ **The theory.** I shall now try to describe the economic theory more precisely and to state what I believe to be its strengths and weaknesses. The theory is based on two simple but important insights. The first is that since the coercive power of government can be used to give valuable benefits to particular individuals or groups, economic regulation—the expression of that power in the economic sphere—can be viewed as a product whose allocation is governed by laws of supply and demand. The second insight is that the theory of cartels may help us locate the demand and supply curves.

Viewing regulation as a product allocated in accordance with basic principles of supply and demand directs attention to factors bearing on the value of regulation to particular individuals or groups, since, other things being equal, we can expect a product to be supplied to those who value it the most. It also directs our attention to the factors bearing on the cost of obtaining regulation. The theory of cartels illuminates both the benefit and the cost side. The theory teaches that the value of cartelization is greater, the less elastic the demand for the industry's product and the more costly, or the slower, new entry into the industry (or cartelized markets within the industry) is. The theory identifies two major costs of cartelization (besides punishment costs, which are relevant only where cartelization is forbidden by law). The first is the cost to the sellers of arriving at an agreement on the price to be charged by and the output of each seller. This agreement determines the profits of each cartel member. The second cost is the cost of enforcing the cartel agreement against nonparticipants or defectors. Cartels are plagued by "free rider" problems. After the sellers agree to charge the price that maximizes their joint profits, each seller has an incentive to sell at a slightly lower price, because his profits are likely to be higher at the much greater sales volume that a slightly lower price will enable him to obtain. If enough sellers submit to the temptation, the cartel will collapse. A cartel is particularly fragile if members are able to conceal price cuts from one another; then each has the hope of being able to obtain substantial short-term profits before the other members realize that he is cutting price and match him.[23]

Since the effect of typical regulatory devices (entry control, minimum rates, exemption from the antitrust laws) is the same as that of cartelization—to raise prices above competitive levels—the benefit side of cartel theory is clearly relevant. The cost side also seems relevant. The members of the industry must agree on the form of regulation. And just as the individual seller's profits are maximized if he remains outside of the cartel (as long as his competitors remain inside), so any individual or firm that would be benefitted by a type of regulation will have some incentive to avoid joining in the efforts of his group to obtain the regulation. If the regulation is forthcoming, he will benefit from it—he cannot be excluded from the protection of a general regulation, just as a seller cannot be excluded from the benefits of his competitors' charging a monopoly price—but, unlike the active participants in the coalition, he will benefit at no cost.

---

[22] See Truman [44], p. 512.

[23] For a summary of the theory of cartels, see Posner [36].

The theory of cartels teaches that the reluctance to cooperate in maintaining a monopoly price is most likely to be overcome if the number of sellers whose actions must be coordinated is small, which tends to reduce the costs of coordination and of policing, and if the interests of the sellers are identical or nearly so, which should reduce the cost of securing agreement.[24] Likewise in the regulatory sphere, the fewer the prospective beneficiaries of a regulation, the easier it will be for them to coordinate their efforts to obtain the regulation. Also, it will be more difficult for one of them to refuse to participate in the cooperative effort without causing the effort to collapse. Thus, all will tend to participate, knowing that any defection is likely to be followed promptly by the defection of the remaining members of the group, leaving the original defector worse off than if he had not cooperated. The homogeneity of the interests of the members is also significant. The more homogeneous their interest in the regulation in question, the easier (cheaper) will it be for them to arrive at a common position and the more likely will it be that the common position does not so disadvantage one or more members as to cause them to defect from the group.[25]

The analysis of cartels is plainly relevant to the development of an economic theory of regulation, but it is not that theory. If it were, we would observe the same industries obtaining regulatory protection as form durable cartels. We do not. Many industries, such as agriculture, certain occupations, many branches of retail trade, and some manufacturing industries such as textiles, which have obtained favorable regulation, lack the characteristics that predispose a market to cartelization, in particular fewness of sellers. Casual observation suggests that highly concentrated industries are actually *less* likely to obtain favorable regulation than less concentrated industries,[26] reversing the usual expectation with regard to the incidence of cartelization.

There are two reasons why the pattern of regulation and the pattern of private cartelization are different. First, the demand for regulation (derived from its value in enhancing the profits of the regulated firms) is greater among industries for which private cartelization is an unfeasible or very costly alternative—industries that lack high concentration and other characteristics favorable to cartelizing. They lack good substitutes for regulation. (This point suggests, incidentally, a testable—in principle anyway—hypothesis of the economic theory of regulation: among randomly selected unconcentrated industries the presence of cartel-like regulation will be negatively related to the price elasticity of demand for the industry's product *at the competitive price.* The qualification, which is critical, makes the test difficult to carry out in practice.)

---

[24] The characteristics that predispose a market to cartelization are discussed in Posner [34], pp. 116–117.

[25] On the other hand, the more successful and profitable the cartel, the greater the costs to consumers, and so the greater the incentive of consumers to organize against the cartel. Stigler has suggested that the role of the "outsider" (e.g., the consumer) is greater in the public regulation than in the private cartelization context (see [42], p. 16), but it is not clear why a cohesive group of customers would not be equally effective in exacting concessions from a private cartel.

[26] For some evidence in support of this hunch, see MacPherson [26].

Second, whereas cartelization is the product purely of the cooperative action of the firms, favorable regulation requires, in addition, the intervention of the political process. Some industries may be able to influence that process at lower cost than others and these may not be the same industries that are able to cartelize at low cost. In particular, the political dimension of regulation requires two modifications of the theory of cartels as applied to regulation. First, as Stigler proposes in his paper on the free-rider problem, which appears as an appendix to this article, each member of an industry will have an interest in participating in the coalition seeking protective regulation when there is significant asymmetry among the positions of industry members. Protective regulation can take a variety (greater than in the case of private cartelization) of forms—limitation of entry, cash subsidy, tariff, etc.—and the choice of the form may, assuming asymmetry among the positions of the industry's members, affect differentially the welfare of those members. If so, each will want to participate in the industry campaign for regulation so that the choice of the form of regulation to seek will reflect his views. The free-rider problem will still be easiest to overcome where the number of firms in an industry is small, but if the asymmetry condition is fulfilled, even the presence of many firms may not erect an insurmountable obstacle to the formation of an effective coalition. This suggests that it may be cheaper for large-number industries to obtain public regulation than to cartelize privately.

Second, the determinants of political influence must be worked into the supply side of the market in regulation. But before this can be done it is necessary to specify the character of the political system under discussion: the political system of the Soviet Union—or of the City of Chicago—is not identical to that of the United States.

One can distinguish three distinct forms of political system, all of which play some role in the actual political systems of democratic countries such as the United States. One system I shall call "entrepreneurial:" favorable legislation is sold[27] to the industries that value it most. For the reason just mentioned, these would not be the same industries that form private cartels. The costs of cooperative action are irrelevant under this system: the government can use its taxing or other powers of coercion to enable the industry to overcome any free-rider problem it might have, in order that the industry can raise the maximum purchase price for the legislation.

The next system to be considered is the "coercive:" legislation is awarded to groups that are able to make credible threats to retaliate with violence (or disorder, or work stoppages, or grumbling) if society does not give them favorable treatment. We lack good theories of threats or violence but as a first approximation it would seem that the number of people in the group would be an important determinant of its ability to make credible threats of *serious* disorder or violence (as opposed to threats of minor sabotage, annoying and costly but not deeply threatening).

---

[27] At what price? The government has a monopoly of the sale of regulation so presumably it will be able to charge a positive price even if the cost of supplying regulation is zero. In fact the cost is greater than zero, both because the production of legislation is costly (see text at note 12, *supra*) and because regulation that favors one group imposes costs on others.

The third system is the "democratic:" legislation is awarded by the vote of elected representatives of the people. This system, like the coercive, emphasizes the importance of numbers: not of threateners but of voters. The groups are not identical, but there is great overlap, so we are led to predict that the economic legislation of dictatorial regimes will broadly resemble that of democratic ones—as seems on casual observation to be the case. Willingness to pay is also important in the democratic as in the enterpreneurial political system, since legislators are elected in campaigns in which the amount of money expended on behalf of a candidate exerts great influence on the outcome. However, unlike the case of an entrepreneurial system, in a democratic system the free-rider problem remains a serious one: it may limit the ability of an industry or other interest group to make substantial campaign contributions.

The foregoing analysis suggests that while the characteristics that predispose an industry to successful cartelization may also help it to obtain favorable government regulation, one characteristic that discourages cartelization—a large number of parties whose cooperation is necessary to create and maintain the cartel—encourages regulation. Large numbers have voting (and, potentially, coercive) power and also increase the likelihood of an asymmetry of interests that will encourage broad participation in the coalition seeking regulation. In addition, large numbers, and other factors that discourage private cartelization, increase the demand for protective legislation.

The economic theory can thus be used to explain why we so often observe protective legislation in areas like agriculture, labor, and the professions, where private cartelization would hardly be feasible. This is an important advance over the other theories that we have examined. However, the economic theory has not been refined to the point where it enables us to predict specific industries in which regulation will be found. That is because the theory does not tell us what (under various conditions) is the number of members of a coalition that maximizes the likelihood of regulation. Formally, this is the number beyond which the loss of group cohesiveness caused by adding another member would outweigh the increase in the feasibility and attractiveness of becoming regulated produced by greater voting power and by greater demand for regulation due to greater difficulty of cartelizing privately.

I used to think that there was one case in which the theory yielded an unequivocal and testable prediction. That is where the number of *firms* in the industry is small, thereby facilitating the organization of the industry for effective political action,[28] but the number of *employees* in the industry is great.[29] Since the profits from protective regulation can be divided between the employees and the firms through collective bargaining, it should be possible for the firms to induce the employees to "lend" their voting power to obtaining such regulation. The industry does not quite have the best of both worlds, because the firms' profits from favorable legislation, and hence their incentive to seek it, will be diminished by the amount of the payoff to the employees. This may be considerable. Legislation favorable to the industry, by raising prices, will reduce output and hence the industry's demand for inputs, including labor. The reduction in demand will harm not only the employees who are laid off but the remaining employees as well, since the diminution in the number of employees will reduce their voting power, which they might want to exercise in other areas. These costs will presumably be considered by the union when it negotiates for its share of the profits conferred by the regulation being sought by the firms.

---

[28] I assume that the free-rider problem is least serious when the number

The major problem with this hypothesis is that the small number of firms is a factor that, by reducing the costs of private collusion, reduces the industry's demand for favorable legislation. So the economic theory is not refuted by observing that the most conspicuous example of such an industry—the automobile industry—seems to have been unsuccessful either in obtaining favorable regulation or in warding off unfavorable regulation (such as safety and emission controls). Anyway, the automobile example—like so many in this field—is ambiguous.[30]

As this example suggests, the economic theory is still so spongy that virtually any observations can be reconciled with it. Consider, as a further example, the apparent paradox that so many regulated industries appear to be either extremely atomistic (like agriculture) or extremely concentrated (like local telephone or electrical service). The former would appear to encounter substantial free-rider problems in organizing a politically effective group; the latter would appear to have little demand for regulation. The moderately con-

---

of parties is very few, albeit larger coalitions might also be able to overcome the problem.

[29] The fundamental distinction between number of firms and number of voters undermines Stigler's hypothesis ([42], p. 7) that small firms will enjoy disproportionate political influence. If the number of employees is proportional to sales, it is not obvious why small firms should be any more important in obtaining favorable regulation than in the formation of a private cartel.

[30] Conceivably safety and emission controls hurt foreign manufacturers more than domestic ones. Without evidence, I find this suggestion somewhat implausible, however. For example, the emission controls reduce engine performance—always a big selling point for American cars—which would seem to hurt the domestic manufacturer more. Also, the argument ignores the fact that many imported cars are manufactured by foreign subsidiaries of domestic manufacturers. The foreign entanglements of the domestic companies may explain, however, why the industry does not enjoy tariff protection. Also to be considered is the fact that while the gasoline tax would seem to reduce the demand for automobiles, the proceeds of the tax are largely earmarked for highway construction—and highways are complementary to automobiles—so the tax may have little adverse effect on the industry after all. The need for further research in this area is dramatically apparent.

centrated industry would seem to have the optimal structure in terms of the costs of obtaining legislation and the benefits to be derived from it. But theory can worm its way out of this hole, too. For the small-number case, we can point out: (1) even a naturally monopolistic industry would gain from legislation that increased the demand for its product (e.g., by suppressing substitutes) or prevented entry;[31] (2) even if the members of the regulated industry do not gain from regulation, other groups, for example groups of customers, may;[32] and (3) concentration or monopoly may itself be the result of regulation. In the large-number case, we can point out that the reluctance of each member of a coalition to participate substantially in it may be dominated by the number of members who participate, albeit very modestly. Is industry *X*, having 10 members, likely to spend more money on trade association activities than industry *Y*, which differs only in that it has 10,000 members? Free-rider problems are presumably not serious in the case of industry *X*. Let us assume that each member of that industry contributes $1,000 for a total of $10,000 and that this approximates the optimal expenditure for the industry. Free-rider problems may be serious in industry *Y*, so serious that it would be impossible for the industry to raise $1,000 from each member were that necessary to reach an optimal level of expenditures. But the industry does not have to raise that amount from each member in order to match industry *X*—to do that it need only raise $1 from each member.

As part of the search for a harder-edged theory of regulation, it has been suggested that the geographic concentration of the people who would benefit from favorable regulation is an important element since a legislator will exert greater efforts on behalf of a voter bloc large enough to influence the outcome of an election materially. But it has not been demonstrated that this is a generally valid proposition. If the same number of voters are more widely dispersed, no legislator will pay as much attention to their demands, but more legislators will pay some attention, and the net effectiveness of the interest group in the legislature *may* (it is an empirical question whether it *will*) be greater. The proposition also ignores the importance of the President in the legislative process. A Presidential candidate has little reason to respond to the desires of voter blocs concentrated in states in which the vote is not expected to be close. Thus we are at a loss to say whether observing a geographically concentrated—or dispersed—group obtaining—or failing to obtain—regulation confirms or refutes the economic theory of regulation. And this illustrates the essential deficiency of the economic theory of regulation in its present form. At best it is a list of criteria relevant to predicting whether an industry will obtain favorable legislation. It is not a coherent theory yielding unambiguous and therefore testable hypotheses.

Another sort of weakness is that the theory, pushed to its logical extreme, becomes rather incredible, because it excludes the possibility that a society concerned with the ability of interest groups to manipulate the political process in their favor might establish institutions that enabled genuine public interest considerations to influence the formation of policy. One can certainly argue that the U.S. Constitution, in establishing an independent judiciary, did just this (and this point is discussed further below). The constitutional requirement of payment of compensation in eminent domain cases is a similar example.[33] More generally, the many features of law and public policy designed to maintain a market system are more plausibly explained by reference to a broad social interest in efficiency than by reference to the designs of narrow interest groups.[34] One can of course say that on some issues the relevant interest group consists of everyone, or almost everyone, in the society. But this usage robs the interest group concept of its utility by collapsing it into the public interest theory.

---

[31] Even under conditions of natural monopoly, the profit-maximizing monopoly price will induce entry, albeit of firms having higher costs than the monopolist.

[32] See Posner [40].

☐ **The evidence.** Let us turn now to the empirical evidence bearing on the economic theory of regulation. There are a fair number of case studies—of trucking, airlines, railroads, and many other industries—that support the view that economic regulation is better explained as a product supplied to interest groups than as an expression of the social interest in efficiency or justice.[35] I shall discuss in a moment the question just how much support for the economic theory of regulation do these studies really provide. But first I want to discuss another type of empirical evidence, so far largely neglected, that provides additional support for the economic interest group approach. This is evidence concerning the procedures employed in the regulatory process.

A corollary of the economic theory of regulation is that the regulatory process can be expected to operate with reasonable efficiency to achieve its ends. The ends are the product of the struggle between interest groups, but, as suggested earlier, it would be contrary to the usual assumptions of economics to argue that wasteful or inappropriate means would be chosen to achieve those ends. We saw that the evidence traditionally adduced to show that regulatory agencies are inefficient is highly ambiguous. I want to go beyond that evidence and note some general features of the regulatory process that suggest it is well designed to achieve the ends posited by the economic theory of regulation.

One is the delegation of regulatory authority by legislatures to administrative agencies. As mentioned earlier, legislatures cannot continuously regulate a complex area; they must delegate much of the regulatory function either to the courts or to administrative agencies. In the area of economic regulation the legislative choice has generally been the administrative agency rather than the court. Lawyers defend this choice on the ground that the public interest purposes assumed to lie behind the legislation can be achieved more efficiently due to (1) the agency's specialization and (2) its independence from political control. The first reason seems specious.

---

[33] See Posner [34], p. 22, note 2.

[34] The role of legal institutions in supporting the market system is a major theme in Posner [34].

[35] See references in note 3, *supra*.

Courts have long handled highly complex economic questions, such as those which arise in antitrust cases, no less efficiently (or more inefficiently) than the agencies. Is a merger case tried before a federal district court likelier to be mishandled than one tried before the FTC, or the ICC?[36] The second reason is illogical. The choice is not between agency and direct legislative regulation—the latter is assumed to be impracticable. The choice is between agency and court, and the court is more insulated from political control than the agency. The terminal character of many judicial appointments, the general jurisdiction of most courts, the procedural characteristics of the judicial process, and the freedom of judges from close annual supervision by appropriations committees, all operate to make the courts freer from the interest group pressures operating through the legislative process, and more disposed to decide issues of policy on grounds of efficiency, than any other institution of government—specifically the administrative agency, where these features are absent or attenuated.[37] If I am correct in suggesting that the judicial process is designed to resist interest group pressures, it would seem to follow that the delegation phenomenon should count as evidence in support of the interest group theory of regulation.

My article, "Taxation by Regulation,"[38] presented some additional evidence of the influence of interest group pressures on the structure and procedures (as distinct from the substantive outcomes) of the regulatory process. The article suggests that a number of standard features of public utility and common carrier regulation, including controls over construction of new plant and over abandonment of service, the duty of the common carrier to serve all comers, and the tendency to impose public utility and common carrier controls on industries that sell services rather than goods, are best explained on the theory that regulation is designed in significant part to confer benefits on politically effective customer groups. Much regulation, I argued, may be the product of coalitions between the regulated industry and customer groups, the former obtaining some monopoly profits from regulation, the latter obtaining lower prices (or better service) than they would in an unregulated market—all at the expense of unorganized, mostly consumer, groups.[39]

Since that article was written, an example has occurred to me where regulation may be the product of an alliance between the industry and a supplier group. A perplexing feature of airline regulation is that although the CAB has evidently been effective in facilitating cartel pricing by the airlines, it has (until very recently[40]) exercised no control whatever over nonprice competition. The effect of unrestricted nonprice competition when price competition is constrained is to increase the costs of the competing firms and thereby reduce their profits, but, under plausible as-assumptions, by less than if they competed in price.[41] It seems that the airline industry has incurred additional costs largely from equipment purchases. The airlines compete with one another by purchasing newer and more comfortable aircraft and by offering more flights and therefore greater convenience to travelers. The airlines may have purchased *more* equipment than they would if they were competing in price as well as in service (although a possibly offsetting fact is that the demand for air travel is less than it would be if the industry's prices were lower). If so, this would suggest that an apparently inexplicable omission in the regulatory scheme may actually be the calculated result of a coalition of interest groups.

---

[36] For some evidence that it is not, see Posner [39].

[37] See Posner [34], chapters 23, 27. An interesting point here is the traditional reluctance of the courts to permit groups to litigate, which is manifested in requirements of "standing" to sue and in prohibitions against "lay intermediaries" between client and lawyer. A trade association cannot bring a lawsuit seeking a legal rule favorable to its members. The member must sue on his own behalf. This reduces the influence of interest groups in the litigation process. But see *NAACP* v. *Button*, 371 U.S. 415 (1963).

[38] [40].

[39] This extension of the economic theory of regulation helps explain, for example, why the original Interstate Commerce Act was supported by (some) shippers as well as the railroads themselves: the railroads' discriminatory pricing—the target of the Act—both undermined the railroads' cartels and harmed shippers competing with favored purchasers.

The body of empirical evidence supporting the economic theory of regulation has, however, several shortcomings.

(1) Most of the evidence is consistent with *any version* of the interest group theory. The evidence relating to the internal efficiency of regulatory agencies does not enable one to discriminate among any specific such theories (such as the economic theory), because none asserts that regulatory agencies are inept.[42] Only the public interest theory is damaged by such evidence. The case studies on the substance of regulatory policy suffer from the same inadequacy. To show that the Interstate Commerce Act was enacted to benefit the railroads, or the Civil Aeronautics Act the airlines, or that the licensure of physicians benefits them rather than their patients, or that much regulation seems subservient to special-interest customer groups, is to show only that interest groups influence public policy. For these case studies to support the economic theory of regulation they would have to demonstrate that the characteristics and circumstances of the interest groups were such that the economic theory would have predicted that they, and not some other groups, would obtain the regulation that we observe them enjoying. Otherwise *any* legislation that benefitted some group at the expense of the general public would count as support for the economic theory of regulation.

I am aware of only three studies that have tried to test the economic theory of regulation, as distinct from the general interest group theory: two by Stigler (of highway weight limitations for trucks and of occupational licensure)[43] and one by McPherson (of tariffs).[44] Only the results of one of the studies (trucks) clearly support the theory.[45]

---

[40] When it approved capacity-limitation agreements among airlines in certain markets.

[41] See Stigler [41].

[42] The theory, discussed in Section 3, that regulatory agencies eventually knuckle under to the regulated firms comes close to implying that regulatory agencies are ineptly managed (otherwise they would not be so easily conquered). This is an example of the confusing overlap, noted earlier, between this version of the capture theory and the public interest theory.

(2) The empirical research has not been systematic. The researcher does not draw a random sample of, say, the economic legislation passed in the last ten years and ask how much of that legislation can be explained by the economic theory of regulation. Instead, he picks the cases that seem from a distance to support the theory[46] and seeks to determine whether that initial impression was correct. I am not criticizing these studies. Had they shown that trucking, and airline, and railroad regulation could *not* be explained by reference to the operation of interest groups, the significance for scholarship would have been immense. But even a lengthy series of case studies cannot provide much support for the economic theory of regulation, given that the industries studied do not appear to be—and were not selected as—typical and that apparent counterexamples abound. The "consumerist" measures of the last few years—truth in lending and in packaging, automobile safety and emission controls, other pollution and safety regulations, the aggressiveness recently displayed by the previously lethargic Federal Trade Commission—are not an obvious product of interest group pressures,[47] and the proponents of the economic theory of regulation have thus far largely ignored such measures. Nor have there been case studies of industries that fail (or never try) to obtain favorable regulation. Furthermore, there is a serious question whether it is proper to define the subject of study as "economic" regulation. Criminal laws, civil rights legislation, legislative reapportionment, and other "noneconomic" regulations affect economic welfare no less than the conventional forms of economic regulation, and it seems arbitrary to exclude them from the analysis: presumably they obey the same laws of social behavior that we think explain economic regulation.

(3) Some of the case studies of regulation have produced evidence difficult to reconcile with the economic theory. I refer in particular to studies which indicate that maximum-price regulation has little or no effect on the price levels of public utilities[48] and that some forms of regulation generate costs in resource misallocation that seem large in relation to the benefits to the favored interest group.[49] Both sorts of evidence may seem to confirm the influence of interest groups in the regulatory process but it is only the crudest form of interest group analysis that they support. There is no basis

---

[43] In [41, 42].

[44] In [26].

[45] Stigler's tests of occupational licensure produced mixed results. A partial test of the theory is also attempted in Pashigian [29], again without successful results.

[46] However, this does not appear to be true of Stigler's study of state limitations on truck weights.

[47] See Peltzman's recent study of the regulation of new drugs by the Food and Drug Administration in [31]; cf. Posner [37].

[48] See Moore [27], Jackson [18], and Stigler and Friedland [43].

[49] The oil import quota program is a notable example. See Cabinet Task Force on Oil Import Control [6], pp. 28–30.

in the economic theory of regulation for ineffectual regulation—for trying and failing to limit the prices of the regulated firms. The obvious explanation is that maximum price controls are a fig leaf which the regulatory agency dons to conceal from the public its domination by an interest group. But the economic theory of regulation—as thus far developed—does not predict that regulatory agencies will practice fraud on the general public.[50]

Nor does the theory predict that legislatures will choose unnecessarily expensive methods of conferring benefits upon effective political groups. Perhaps they do not. It has been estimated that hundreds of millions of dollars a year could have been saved had oil companies received outright grants from the Treasury rather than oil import quotas which, in the process of enriching the companies and the owners of domestic oil-producing property, induce consumers to make inefficient substitutions for oil.[51] But the underlying assumption—that there is a large avoidable deadweight loss—may well be incorrect. An increase in income tax rates to finance an outright grant to the oil companies could have costly substitution effects (e.g., leisure for work) of its own. If the assumption *is* correct, the implications for the economic theory of regulation are disturbing. It is in everyone's interest to use a more rather than a less efficient way of transferring money to the oil companies. Stigler, in his search for a rational explanation of the quotas, has argued that it would be impracticable to give money to the oil companies directly, because then firms would have an incentive to create oil-company affiliates in order to be entitled to the subsidy.[52] However, that danger could be averted by limiting the subsidy to oil companies in existence as of the date of the grant. The CAB gave cash subsidies to the airlines for many years: its control over entry prevented the subsidies from attracting new entrants. Nor are entry controls strictly necessary: the cash grant can be limited to the firms in the industry at the date of the grant (or some earlier date to prevent entry in anticipation of the grant).

(4) The empirical evidence depends heavily on a confident rejection of the public interest rationales in which all legislation is—for reasons not yet illuminated by the economic theory of regulation—cloaked. Sometimes these rationales have just enough plausibility to make such rejection questionable. The oil import quota case is again an interesting one. The recent Arab oil embargo suggests that it is not palpably absurd to adopt governmental policies designed to reduce U.S. dependence on the oil produced by the Arabs. Stigler has argued that if this were the actual purpose behind the oil import quota system, it would have been carried out not by a quota system but by a tariff, since the revenues generated by a tariff would go to the taxpayers rather than to the oil com-

---

[50] Such a prediction would be a logical extension of Stigler's remarks on the relevance of information costs in the analysis of the political process (in [42], pp. 11–12), but the extension has not been made. On the economics of fraud, see Posner [38], pt. 1, and references cited therein. The possible application of the economic theory of fraud to the theory of regulation is discussed in the text below.

[51] See [6].

[52] See Stigler [42], pp. 4–5.

panies' stockholders.[53] (Another alternative would be stockpiling imported oil.) But the argument proves only that the purpose behind the system may have been a mixture of public interest considerations and interest group pressures.[54]

(5) The effects of economic regulation are difficult to trace. A tax on gasoline might help the railroad industry. The cartelization of the airline industry under the CAB's aegis benefits surface transportation (the demand for which is increased by anything that increases the price of a substitute service). These complications make it difficult to identify the industries that benefit from and those that are injured by regulation. It is superficial to point to an industry as an example of an effective political group because it enjoys a high tariff without considering the impact on it of other governmental policies, including many ostensibly imposed on different industries. We do not know whether to regard automobile emission controls as a sign of the industry's inability to ward off adverse regulation or as a token of how limited, and late, government regulation of the automobile industry has been.

(6) An important, but as yet unexplained, datum is the characteristic public interest rhetoric in which discussions of public policy are conducted and the policies themselves framed. The use of language that, if the economic theory of regulation is correct, is utterly uninformative and indeed misleading is not costless; presumably it is employed only because there are offsetting benefits. These benefits must have to do with increasing the costs to members of the public of obtaining accurate information about the effect of the actions of their legislative representatives on their welfare.

Recent developments in the economic theory of fraud may prove helpful in explaining the prevalence of misleading rhetoric in discussions of public policy. The propensity to engage in fraud seems to be related to such factors as the difficulty (cost) of the buyer's determining the performance characteristics of the product (by inspection, use, or whatever) and the value of the buyer's time. The greater the cost of determining the product's performance characteristics, or of the time spent by the buyer in trying to ascertain those characteristics, the more fraud we can expect to find. Where the product is legislation, the cost of determining its quality is often extremely high. With respect to the value of the buyer's time, it is important to note that a legislative proposal must be "sold" to two groups: the legislators and the electorate. Our earlier discussion of the costs of legislation implied that the cost of a legislator's time is very high, which in turn implies that the amount of time he can efficiently devote to appraising the merits of proposed legislation is small.

The introduction of considerations based on the economic analysis of fraud, or more broadly of the costs of information, suggests that it may be possible to revive the public-interest-miscarried theory of regulation in a form that it can be made rigorous and empirically testable.

---

[53] *Id.* p. 4.

[54] Another consideration is that the oil import quota program could be and was established by Executive Order rather than by statute, whereas imposition of a tariff would have required congressional action.

5. Conclusion

■ This article has offered a number of criticisms of both the traditional public interest theory of regulation and the newer economic theory which conceives regulation as a service supplied to effective political interest groups. Neither theory can be said to have, as yet, substantial empirical support. Indeed, neither theory has been refined to the point where it can generate hypotheses sufficiently precise to be verified empirically. However, the success of economic theory in illuminating other areas of nonmarket behavior leads one to be somewhat optimistic that the economic theory will eventually jell: the general assumption of economics that human behavior can best be understood as the response of rational self-interested beings to their environment must have extensive application to the political process.

References

1. BAXTER, W. F. "NYSE Fixed Commission Rates: A Private Cartel Goes Public." *Stanford Law Review*, Vol. 22 (April 1970), pp. 675–712.
2. BENTLEY, A. F. *The Process of Government*. Chicago: University of Chicago Press, 1908.
3. BERNSTEIN, M. *Regulating Business by Independent Commission*. Princeton: Princeton Univ. Press, 1955.
4. BONBRIGHT, J. C. *Principles of Public Utility Rates*. New York: Columbia Univ. Press, 1961.
5. BUCHANAN, J. AND TULLOCK, G. *The Calculus of Consent*. Ann Arbor: University of Michigan Press, 1962.
6. CABINET TASK FORCE ON OIL IMPORT CONTROL. *The Oil Import Question*. Washington, D.C.: U.S. Government Printing Office, 1970.
7. COASE, R. H. "The Federal Communications Commission." *Journal of Law and Economics*, Vol. 2, No. 2 (October 1959), pp. 1–40.
8. COMANOR, W. S. AND MITCHELL, B. "The Costs of Planning: The FCC and Cable Television." *Journal of Law and Economics*, Vol. 15, No. 1 (April 1972), pp. 177–231.
9. DAVIS, K. C. *Administrative Law Treatise*. St. Paul: West, 1958 and 1970 supplement.
10. DOWNS, A. *An Economic Theory of Democracy*. New York: Harper, 1957.
11. ECKERT, R. D. "What Do Regulatory Commissions Maximize?" Unpublished manuscript, University of Southern California, 1972.
12. EHRLICH, I. AND POSNER, R. A. "An Economic Analysis of Legal Rulemaking." *Journal of Legal Studies*. Vol. 3, No. 1 (January 1974), pp. 257–286.
13. FRIENDLY, H. J. *The Federal Administrative Agencies: The Need for Better Definition of Standards*. Cambridge: Harvard Univ. Press, 1962.
14. HERRING, E. P. *Public Administration and the Public Interest*. New York: McGraw-Hill, 1936.
15. HILTON, G. W. "The Consistency of the Interstate Commerce Act." *Journal of Law and Economics*, Vol. 9, No. 2 (October 1966), pp. 87–113.
16. HIRSHLEIFER, J., DEHAVEN, J. C., AND MILLIMAN, J. W. *Water Supply: Economics, Technology, and Policy*. Chicago: University of Chicago Press, 1960.

17. HUNTINGTON, S. P. "The Marasmus of the ICC: The Commission, the Railroads, and the Public Interest" in P. Woll, ed., *Public Administration and Policy: Selected Essays,* New York: Harper & Row, 1966.
18. JACKSON, R. "Regulation and Electric Utility Rate Levels." *Land Economics,* Vol. 45, No. 3 (August 1969), pp. 372–376.
19. JORDON, W. A. *Airline Regulation in America: Effects and Imperfections.* Baltimore: Johns Hopkins Press, 1970.
20. ———. "Producer Protection, Prior Market Structure and the Effects of Government Regulation." *Journal of Law and Economics,* Vol. 15, No. 1 (April 1972), pp. 151–176.
21. KITCH, E. W., ISAACSON, M., AND KASPER, D. "The Regulation of Taxicabs in Chicago." *Journal of Law and Economics,* Vol. 14, No. 2 (October 1971), pp. 285–350.
22. KOLKO, G. *Railroads and Regulation, 1877–1916.* Princeton: Princeton Univ. Press, 1965.
23. LEISERSON, A. "Interest Groups in Administration" in F. Morstein Marx, ed., *Elements of Public Administration,* New York: Prentice-Hall, 1946.
24. MACAVOY, P. W. *The Economic Effects of Regulation: The Trunk-Line Railroad Cartels and the Interstate Commerce Commission before 1900.* Cambridge: M.I.T. Press, 1965.
25. ———. "The Regulation-Induced Shortage of Natural Gas." *Journal of Law and Economics,* Vol. 14, No. 1 (April 1971), pp. 167–199.
26. MACPHERSON, C. B. "Tariff Structures and Political Exchange." Unpublished Ph.D. dissertation, University of Chicago, 1972.
27. MOORE, T. G. "The Effectiveness of Regulation of Electric Utility Prices." *Southern Economic Journal,* Vol. 36, No. 4 (April 1970), pp. 365–375.
28. OLSON, M., JR. *The Logic of Collective Action.* Cambridge: Harvard Univ. Press, 1965.
29. PASHIGIAN, B. P. "Public vs. Private Ownership: Consequences and Determinants of Local Transit Systems." Unpublished manuscript, University of Chicago Graduate School of Business, October 1973.
30. PELTZMAN, S. "Entry in Commercial Banking." *Journal of Law and Economics,* Vol. 8, No. 2 (October 1966), pp. 11–50.
31. ———. "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments." *Journal of Political Economy,* Vol. 81, No. 5 (September–October 1973), pp. 1049–1091.
32. POSNER, R. A. "The Appropriate Scope of Regulation in the Cable Television Industry." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 98–129.
33. ———. "The Behavior of Administrative Agencies." *Journal of Legal Studies,* Vol. 1, No. 2 (June 1972), pp. 305–323.
34. ———. *Economic Analysis of Law.* New York: Little, Brown and Co., 1973.
35. ———. "Natural Monopoly and Its Regulation." *Stanford Law Review,* Vol. 21 (February 1969), pp. 548–643.
36. ———. "Oligopoly and the Antitrust Laws: A Suggested Approach." *Stanford Law Review,* Vol. 21 (June 1969), pp. 1562–1575.
37. ———. "Reflections on Consumerism." *University of Chicago Law School Records,* Vol. 20, No. 3 (Spring 1973), pp. 19–25.
38. ———. *Regulation of Advertising by the FTC.* American Enterprise Institute, November 1973.
39. ———. "A Statistical Study of Antitrust Enforcement." *Journal of Law and Economics,* Vol. 13, No. 2 (October 1970), pp. 365–419.
40. ———. "Taxation by Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 22–50.
41. STIGLER, G. J. "Price and Nonprice Competition" in G. J. Stigler, ed., *Organization of Industry,* Homewood, Ill.: Richard D. Irwin, 1968.
42. ———. "The Theory of Economic Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 3–21.

43. ——— and Friedland, C. "What Can Regulators Regulate? The Case of Electricity." *Journal of Law and Economics,* Vol. 5, No. 2 (October 1962), pp. 1–16.
44. Truman, D. B. *The Government Process: Political Interests and Public Opinion.* New York: Knopf, 1951.
45. Wilson, J. *The Politics of Regulation.* Washington, D.C.: The Brookings Institution, forthcoming.
46. Ziegler, H. *Interest Groups in American Society.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Free riders and collective action: an appendix to theories of economic regulation

George J. Stigler

Professor of Economics
University of Chicago

The free rider problem is restated more precisely as the cheap rider problem. It is argued that if one takes account of the frequent or typical asymmetry in the interests of different enterprises in an industry, the individual incentives of many enterprises to participate in joint ventures are substantial.

1. Introduction

The free rider proposition asserts that in a wide range of situations, individuals will fail to participate in collectively profitable activities in the absence of coercion or individually appropriable inducements. The proposition does not specify the circumstances favorable to collective action, and hence does it explain why there are innumerable operating and presumably not wholly ineffective collective bodies—literally thousands of trade associations, for example. The present essay seeks to improve upon this situation.

2. The proposition

The free rider proposition is easily illustrated. Let the gain to an individual (firm) be equal to G if a collective activity is undertaken. For example, G is the firm's gain from the tariff which could be obtained by an effective industry lobby. The cost of the collective action is C, and there are n identically sized firms. By hypothesis, the joint-action is collectively profitable, so $nG > C$. The individual will refrain from joining the collective action if n is of some appreciable size, on the assumption that the viability of the action does not depend upon his participation. If enough individuals take this position—and at this level of discourse all are symmetrical in position—the collective action will not be taken.

Even at this level of simplicity, it should be apparent that rides, like lunches, cannot be wholly free.[1] If firm i does not join the collective action, it incurs two costs:

(1) The probability that the collective action is undertaken will be reduced, so the expected gain is reduced.
(2) Even if a sufficient number of the remaining individuals join the collective action, they will not pursue the collec-

---

George J. Stigler received the B.B.A. from the University of Washington (1931), the M.B.A. from Northwestern University (1932), and the Ph.D. from the University of Chicago (1938). Professor Stigler's research interests include citation analysis, the economic theory of regulation, and the theory of legislatures. At this time, however, his most immediate concern is for working on his "Treatise," when he is not skirting other arenaceous loci.

[1] I am indebted to Harold Demsetz for the succeeding argument.

tive action on so large a scale as if firm *i* joined: alternatively put, if *k* firms join and each contributes c_k dollars, if *i* also joins, the total contribution—$(k+1)c_{k+1}$—will surely rise.

Even with these costs, *i* may abstain from the collective action, but his "ride" will be cheap, not free.

There is no critical number of individuals for which this argument fails—or holds. Even with $n = 2$, there may be a cheap rider—one of two airlines using a field may erect the sleeve to show wind direction. With *n* large, whatever that may be, it is commonly believed that the proposition is inexorable.

3. The by-product theory

Moore proposed a possible solution to the problem of reconciling the proposition with the existence of numerous collective bodies:

> If an occupation has an association which furnishes certain services to its members which cannot be purchased at a competitive price elsewhere, then the association is in the position of a monopolist: it faces a downward sloping demand curve. The services it furnishes may be journals, meetings, a certificate of membership, information about new developments, advertisements, and so forth. . . . To the extent that the association faces a downward sloping demand curve, it can act as a monopolist to raise funds for lobbying.[2]

This position was presented by Olson in a more elaborate form in *The Logic of Collective Action*,[3] but also without any attempt at empirical testing.

The by-product theory invites an obvious objection: why should the association be able to charge more than the cost of the services which are appropriable as private goods? If an association seeks to add a charge for collective goods, a rival association which undertook no collective actions could undersell it—and there are no barriers to entry into trade associations. Even if the services, such as the collection of information, have great economies of scale and are "natural monopolies," the argument is not affected: a rival association can still bid away the members of the group with a lower price.[4]

4. The small number solution

When the number of individuals in the group is small, the probability that each individual will join the group is increased. The factors we have already cited for believing that outsiders do not get a *free* ride are applicable here. First, the probability of collective action depends more strongly on each firm's participation; and second, the scale of operation of the collective action will be appreciably smaller if fewer individuals join. Olson has labelled this small number theory a "special interest" theory of group action.[5]

The arithmetical rules of participation may be formulated as follows:

---

[2] [2], p. 114.

[3] See [3], especially Chapter VI.

[4] See Demsetz [1].

[5] In [3], pp. 141 ff.

π_p = the probability of collective action if i joins,
π_{np} = the probability of collective action if i abstains,
$G_i(m, e)$ = the expected gain to i if collective action is taken,
m = the number of individuals joining the coalition, and
$e(m)$ = the expenditures per individual who joins.

Then individual i should join if

$$\pi_p\{G(m, e) - e(m)\} > \pi_{np}\, G(m - 1, e + \Delta e)$$

and, as a Taylor series approximation,

$$G(m - 1, e + \Delta e) = G(m, e) - G_m(m, e) + \Delta e G_e(m, e)$$

so

$$(\pi_p - \pi_{np})G(m, e) - \pi_p e(m) - \pi_{np}\{\Delta e G_e(m, e) - G_m(m, e)\} > 0.$$

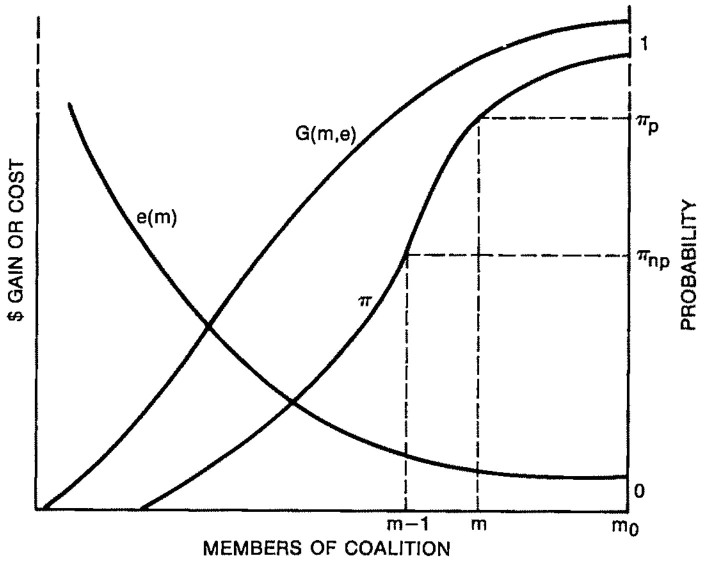
The expected net gain for i from participation is thus:

(1) The increase in the probability of the formation of the coalition times the gain to i if it is formed. The latter term will increase with m; the former probably decreases as m increases.
(2) Minus the expected cost of contributions to the coalition. Since π_p increases with m, and e probably decreases with m (diminishing returns to coalition expenditures), the net effect of changes in m is again uncertain.
(3) Minus the effect on i's gain of the smaller size of the coalition if i does not join,[6] times π_{np}.

Plausible forms of the various functions are displayed in Figure 1, where m_0 is the total number of possible members of the coali-

FIGURE 1

PLAUSIBLE FORMS OF VARIOUS FUNCTIONS



---

[6] $\Delta G = \Delta e G_e + \Delta m G_m = \Delta e G_e - G_m$.

tion. The smaller m_0, the steeper the probability curve (π), the more rapid the fall in $e(m)$, and the greater the rate of rise of $G(m, e)$—all for a given m. Or so the folklore wisdom has it.

The small number solution has a wider scope than a literal count of numbers would suggest. The size distribution of individuals is highly skewed when these individuals have a size dimension (sales of firm, property of family). The large individuals in a group may therefore properly view themselves as members of a small number industry if their aggregate share of the group's resources is large. Consider the example, cigar manufacturers. The 1963 data report:

4 largest firms—66 percent of value added
next 4 firms —17 percent of value added
next 12 firms —10 percent of value added
next 30 firms — 6 percent of value added
next 143 firms— 1 percent of value added.

One can debate the proper way to convert this array into an equivalent number of equal-sized firms. If one uses the Herfindahl measure of concentration the equivalent number is about 7 or 8.[7] Many, many industries fulfill in good measure the small number condition.

5. The asymmetry solution

■ With trivial exception, every industry produces a wide variety of goods or services. The fabrics in textiles or the metals and woods in furniture, the weight and power of automobiles, the composition and mechanical properties of steel—all are found in profuse variety in the products of industry. The smaller firms in an industry seldom make the full range of products: they specialize in a narrower set of products. Hence, if they are not represented in the coalition, they may find that their cheap ride is to a destination they do not favor. The proposed tariff structure may neglect *their* products; the research program may neglect *their* processes; the labor negotiation may ignore *their* special labor mix.

We may fashion a measure of the divergence of interests of the large and small firms. Let:

$s_i =$ the share of sales of commodity i made by the combined large firms,
$1 - s_i =$ the share of sales made by the combined small firms,
$S_i =$ the total sales of commodity i,
$S_L, S_s =$ the total sales of large and small firms, respectively,
$S_L + S_s = S$, total sales.

The relative importance of sales of commodity i to the large firms is

$$\frac{s_i S_i}{S_L},$$

and the difference between large and small firms in this relative importance is

---

[7] See Stigler [4], Chapter 4.

$$\Delta_i = S_i \left(\frac{s_i}{S_L} - \frac{(1 - s_i)}{S_s} \right).$$

If we define the concentration ratio, *k*, as the share of total sales made by large firms,

$$k = \frac{S_L}{S_L + s_s},$$

and

$$\Delta_i = \frac{S_i}{S} \left(\frac{s_i}{k} - \frac{(1 - s_i)}{1 - k} \right).$$

Hence we may use, as our measure of the discordance of interest between large and small firms,

$$\Delta^2 = \Sigma \frac{S_i^2}{S^2} \left(\frac{s_i}{k} - \frac{(1 - s_i)}{(1 - k)} \right)^2$$

$$= \frac{1}{S^2 k^2 (1 - k)^2} \Sigma S_i^2 (s_i - k)^2.$$

Our central explanation for group action is that Δ^2 can usually be made sufficiently large to bring most of the larger firms into the association.

6. A sample of associations

■ Whatever the theory says, there is no doubt of the existence of thousands of trade associations in the United States. Often several associations appear—so far as titles can be trusted—to represent the same industry, and often association titles contain trade jargon not necessarily comprehensible to the outsider. To obtain at least a model sample of trade association data, we took a random sample of 70 4-digit Census of Manufacturers (1967-basis) industries,[8] for each of which we chose two companies, with large plants for which that industry was important, from the 1966 *Fortune Plant and Product Directory.*[9] We addressed inquiries to each of the two companies (often subdivisions of larger enterprises) for the name of the trade association(s) which represented them in their activities in the given industry (named by us). The association was then requested to send the literature on its activities and also to report its budget, the number of professional employees, the number of company members, and the share of the covered industry that was represented in the association.

A different type of sample would have been obtained if instead we had begun with associations and inquired as to the activities of their members. The difference is suggested by the following tabulation of useful responses to our inquiries:

---

[8] Confining the sample to industries with coverage and specialization ratios over 75 percent, other than "n.e.c." industries.

[9] I wish to thank Claire Friedland for performing the statistical work. We actually chose plants producing the largest 5-digit product in the selected 4-digit industry, so that both census 4- and 5-digit data would be applicable.

(1) Essential coincidence of association coverage with 4- or 5-digit industry.	32 associations
(2) Association covers a 3-digit industry of which our reporter falls in a 4- or 5-digit class.	11 associations
(3) Association covers a 2-digit industry of which our reporter falls in a 4- or 5-digit class.	11 associations
(4) Association covers several industries (not all in one 2-digit industry) of which our reporter industry is one.	7 associations
(5) Association covers other industries, of which most commonly our reporter company is a supplier or a negligible part.	16 associations

This tabulation assumes that the reporter belongs to a clearly defined 4- or 5-digit industry, and of course by product variety or corporate ownership it may belong to a much wider flung enterprise. The broadest associations, such as the National Association of Electrical Manufacturers, contain subdivisions which deal with more closely defined industries (e.g., wire and cable, appliances, etc.).

A set of regression equations relating association size to industry characteristics is reported in Table 1. In our larger sample, both staff and budget of the association are well related to the size of the covered industry, but not to concentration ratios, which our foregoing discussion would lead us to expect. There is, indeed, some hint of a negative correlation of association resources and

TABLE 1

TRADE ASSOCIATION REGRESSIONS*

X_1=ASSOCIATION BUDGET, THOUSANDS OF DOLLARS, 1969-1970
X_2=ASSOCIATION STAFF, 1969-1970
X_3=NUMBER OF MEMBERS OF ASSOCIATION, 1969-1970
X_4=VALUE ADDED OF MATCHED INDUSTRY, MILLIONS OF DOLLARS, 1967
X_5=1967 CONCENTRATION RATIO, 4-DIGIT
X_6=SHARE OF INDUSTRY BELONGING TO ASSOCIATION, PERCENT, 1969-1970

	N**	R^2
$X_2=10.44+0.00616X_3+0.00667X_4$ (2.71) (3.02)	60	0.258
$X_2=17.40+0.00566X_3+0.00699X_4-0.191X_5$ (2.46) (3.15) (1.22)	60	0.277
$X_1=554.7+0.255X_3+0.742X_4$ (1.22) (3.67)	60	0.230
$X_1=1148+0.212X_3+0.769X_4-16.28X_5$ (1.00) (3.78) (1.14)	60	0.247
$X_2=-19.37+0.0066X_3+0.0039X_4+0.446X_6$ (1.45) (0.72) (1.51)	42	0.137

*t RATIOS IN PARENTHESES.
**NUMBER OF INDUSTRIES FOR WHICH DATA AVAILABLE.

concentration ratios. Can it be that very large companies undertake many of the activities allocated to associations in less concentrated fields? The number of members seems directly related to association activity, the share of industry coverage much less so. This finding is consistent with expectations of the asymmetry theory.

The crucial omission from our sample is any direct knowledge of the role of asymmetry in the economic interests of members.

References

1. Demsetz, H. "Why Regulate Utilities?" *Journal of Law and Economics,* Vol. 11, No. 2 (April 1968), pp. 55–66.
2. Moore, T. G. "The Purpose of Licensing." *Journal of Law and Economics,* Vol. 4, No. 4 (October 1961), pp. 93–117.
3. Olson, M. *The Logic of Collective Action.* Cambridge: Harvard Univ. Press, 1965.
4. Stigler, G. J. *The Organization of Industry.* Homewood, Ill.: Richard D. Irwin, 1968.

Accident law and resource allocation

Peter A. Diamond

Professor of Economics
Massachusetts Institute of Technology
and Visiting Fellow
Balliol College

This article examines the effects from introducing a "frictionless" legal system into a model of competitive equilibrium with externalities. In a traditional negligence–contributory negligence legal system, there are two important legal parameters under "due care" standards. This paper relates the different equilibria which can occur to the levels of the "due care" standards.

1. Introduction

■ In some externality situations there exists the nonmarket mechanism of civil lawsuits to shift costs between parties. Resource allocation is affected by the shifting of costs. It is also affected by the rules determining how much is to be shifted and when shifting is to take place. Individual decisions that affect the likelihood and severity of several-person accidents are prime examples of externalities subject to civil suit.[1] Where the legal system uses a theory of negligence to determine when costs are to be shifted, a person who has suffered losses from an accident can receive compensation only if two conditions are satisfied. The person who caused the accident must have acted in a way that the court views as negligent. In addition the person trying to collect for his losses must not have contributed to the causation of the accident in a way that is adjudged contributorily negligent. The two standards for due care (i.e. the minimal levels of care to avoid being found negligent or contributorily negligent) can play a key role in the effects of the legal system on resource allocation. A natural question for a student of legal decisions is how general theories of the definition of negligence get applied to particular situations to determine which behavior will be found negligent.[2] A natural question for an economist is how the particular definitions of due care applied to a recurrent accident situation affect the equilibrium values of decision variables. This paper relates the equilibrium values of decisions of potential accident participants to the different due care

---

Peter A. Diamond received the B.A. degree from Yale University in 1960 and the Ph.D. from the Massachusetts Institute of Technology in 1963. At M.I.T. Professor Diamond's research centers on public finance.

The author is indebted to the participants in the Oxford workshop in economic theory for helpful comments and to the National Science Foundation for financial assistance.

[1] Examples of recent analyses of accident law and resource allocation are found in Calabresi [2], Posner [5], Brown [1], and Diamond [3].

[2] Prosser describes negligence as "1. A duty, or obligation, recognized by law, requiring the actor to conform to a certain standard of conduct, for the protection of others against unreasonable risks. 2. A failure on his part to conform to the standard required," in [6], p. 143.

standards which might be set by judicial proceedings or legislative action.

To introduce the accident technology, the paper begins with a simplified conventional model of competitive equilibrium with externalities. It is assumed that individuals are engaged in one of two activities in which they decide on the level of a single control variable, which will be called care. All the participants in an activity are the same. The expected utility of an individual depends on his own choice of care and the care taken by everyone in the other activity. Full knowledge and rationality on the part of all participants are assumed.[3] Therefore, equilibrium is defined as simultaneous expected utility maximizing decisions by all individuals under the assumptions that the levels of care chosen by everyone else are given and correctly perceived.

Into this model we shall then insert a "frictionless" legal system and apply the traditional negligence rules as the only corrective social instrument (i.e., we ignore fines, taxes, or other tools to alter decisions). The legal system has three jobs in this model. It announces the due care standards which will be used in the determinations of negligence.[4] After each accident it compares the care taken by the two parties to the accident to the announced standards and holds the party in activity one liable if his care is less than his due care standard and that of the party in activity two at least as large as his due care standard.[5] If liability is assessed, the court determines how much money is to be transferred between the parties, i.e., it measures damages. (It is assumed that these damages are always paid by the liable party.) Individuals are assumed to know exactly how the legal system works. To fit money transfers into the model conveniently, we assume that everyone is risk neutral, i.e., that utility functions are linear in income available to spend on all other goods.[6] For each pair of due care standards we shall determine what equilibrium (or equilibria) can occur.

The analysis of equilibrium begins with individual decisions under this negligence regime. Depending on the behavior of those in the other activity and the due care standards, individuals in activity one select one of three possible maximizing levels of care —(1) maximizing expected utility ignoring externalities, (2) maximizing expected utility net of the expected costs of those with whom one might have an accident, or (3) selecting a care level precisely equal to the due care standard. The different possibilities arise because expected utility may not be continuous in care, as it may sometimes have a discontinuity when care equals the due care standard. A similar analysis is done for choice in activity two.

---

[3] Full rationality is a poor assumption in situations involving one's own safety. The model should be viewed as a first step, awaiting alternative models of choice, or as applying to situations where accidents involve only property damage or risks to others than the decisionmakers, as when managerial decisions affect accidents between employees of different firms.

[4] It is assumed that everyone knows the due care standards, once announced.

[5] It is assumed that all accidents involve just one person from each activity. Except for a brief discussion in Section 8, we shall not consider the possibility of finding liable a person in activity two.

[6] This assumption is equivalent to the alternative assumption of a perfect insurance market which results in each person's bearing the expected value of the costs he would incur. The interaction of the legal system with imperfect insurance is obviously of great interest.

These supply curves of care are then combined to determine the pairs of due care standards which give rise to each of the four types of uniform equilibria which can occur. (Equilibrium is referred to as uniform when all the participants in an activity select the same care level.) The efficient solution is considered relative to the possible outcomes. Because of the discontinuity in utility, there can be two care levels which maximize utility for an individual. This gives rise to the possibility of nonuniform equilibria where participants in one or both activities, although identical, are selecting different care levels. It is argued that an equilibrium always exists under the assumptions made, so that the regions of the three types of nonuniform equilibria are at least as large as the region of due care standards which do not give rise to uniform equilibria. Also briefly considered is a model where those in activity two cannot fully control their own behavior. Rather, they select the determinant of the probability distribution of behavior which contributes to accidents. However, it is assumed that courts still base liability decisions on behavior at the time of the accident rather than on average behavior or choice of the control variable determining the probability distribution. This change in assumptions significantly changes the nature of the analysis.

2. Legal system

■ Since a frictionless negligence system could be modeled in several ways, let us briefly consider the legal institutions which are being so starkly portrayed.[7] Some due care standards are determined by legislative action, in that standards which are legislated for safety, like speed limits, are taken by the courts to define the due care standard.[8] Other standards get determined over time by applications by juries and judges of the general definition of negligence to recurring factual situations. In time, a standard may become a rule of law and replace the application of the general definition to repeating fact situations. In the former case it is reasonable to see the behavior appropriate to avoid negligence defined in physical units like speed or the quality of safety devices. In the latter case the same view of the legal system may be reasonable in a long-run equilibrium, where the underlying technology does not change.

Given a particular accident, the court needs to determine two facts to evaluate negligence (or contributory negligence). First, was there behavior which did not meet the due care standard? Second, was this behavior a sufficiently important cause of the accident to generate liability when the due care standard was not met? (Behavior which is sufficiently important is called a proximate cause of the accident.) In the model the care decision is always assumed to be a proximate cause of the accidents which are being analyzed.[9] Thus, this range of legal issues is completely ignored.

---

[7] For some alternative approaches, see Brown [1].

[8] See, e.g., Prosser [6], section 6.

In measuring damages after an accident, there are costs which are regularly evaluated by courts, like property damage, medical costs, lost wages, pain and suffering, and other costs which are not measured because they are too small or too speculative in measurement.[10] In the model we shall preserve this distinction among the costs of those in activity two, referring to them as recognized and unrecognized costs.[11] (We shall ignore the situation where the obligation to mitigate accident damages leads to a measurement which differs from the entire recognized costs from the accident.)

The best way to justify the assumption in the model that those in activity one can never collect damages from those in activity two is to assume a factual situation where the only costs for those in activity one are not recognized costs. For example, it is unpleasant for most car drivers to strike or have struck a pedestrian, but these psychic costs are not recognized when there are no other damages. Damage to the automobile or its driver would, of course, be recognized. The need for this approach comes from the theoretical role of due care standards as applying to both negligence and contributory negligence. Thus if someone is contributorily negligent, and so barred from collecting damages, he also may be negligent and so potentially liable for the costs of the other party, when the other party is not contributorily negligent.[12] Also there have been circumstances where some parties are legally or practically exempted from tort liability—for example the government or a worker who damages machinery (and is hurt himself) in a setting where a suit would be disruptive to union-management relations. But these examples are probably not of significance. This simplification can, of course, be viewed merely as a simplification before tackling the more complicated symmetric case, which is introduced in Section 8.

The assumption that a fact situation calling for damage payments means that payment is actually made covers several behavioral assumptions. On the plaintiff's side the case is always carefully brought in good time not to jeopardize the suit. The defendant must be available for suit and financially able to pay the damage award. Where the defendant should win on the facts, it is again assumed that he does. In addition it is assumed that settlements out of court do not differ from the outcomes which this perfect court would generate. I suspect this is no more and no less plausible than a perfect court, and represents a second source of friction (and perhaps bias) in the actual system.[13]

---

[9] My earlier paper, cited above, briefly considers decision variables which are not proximate causes.

[10] "No general agreement has yet been reached as to the liability for negligence resulting in fright, shock, or other 'mental distress' or its physical consequences," Prosser [6], p. 327.

[11] Inaccuracy of court measurement of damages can also result in this distinction between costs. The choice of which costs to recognize might be analyzed as a control variable rather than a parameter.

[12] For the simplified accident settings considered here this description is adequate. For a discussion of the relation between negligence and contributory negligence see Prosser [6], pp. 418–421.

It is also assumed that apart from the tort system, there are no socially generated incentives affecting care decisions. This ignores the presence of fines and other punishments which are triggered by the same behavior that will be found negligent in case of an accident.

3. No-liability equilibrium

■ Let us start by considering equilibrium in the absence of any compensation, i.e., with each person bearing the losses which occur to him as a result of accidents. This will serve to introduce formally the model of accident technology and the particular assumptions being made. The model is a conventional model of competitive equilibrium with externalities. We assume that there are a large number of people engaged in each of two activities (with both numbers normalized to unity). All the participants of any activity are assumed to be identical and the numbers participating are taken as given (i.e., there is no decision of whether to participate in the activity). For simplicity we assume that no one participates in both activities. We define equilibrium to be simultaneous expected utility maximizing decisions by all individuals under the assumptions that the levels of care chosen by everyone else are given and correctly perceived. We shall refer to the equilibrium as uniform when all the participants in each activity select the same level of care.

Let us denote by x and y the levels of care chosen by the typical participant in activities one and two, respectively.[14] (We shall consider nonuniform equilibria separately, below.) We assume that both x and y lie in the unit interval. (We shall make assumptions on the utility functions such that optimizing values of care are always in the interior of the unit interval. This will save us from worrying about corner conditions.) Let $A(x,y)$ be the level of expected utility of someone engaged in activity one who has chosen care level x when everyone in activity two has chosen care level y. A will reflect the utility and financial costs of taking care as well as the expected accident costs. To make sense of the shift of accident costs by the legal system as modeled, we shall assume that individuals are risk neutral in bearing accident costs. Thus we assume that utility is linear in income available to spend on other goods. However, we incorporate expenditures on all other goods within A. We assume that A is strictly concave in x and twice continuously differentiable in both variables. To ensure internal maximization we assume that marginal utility of care is positive when care is zero

$$A_x(0, y) > 0 \qquad \text{for all } y \tag{1}$$

(where subscripts x and y on functions refer to partial derivatives).

---

[13] See, e.g., Leff [4].

[14] Since care does not have natural units (and is not purchased or sold), we assume that there is some method of measuring care which results in the satisfaction of all the assumptions made below.

Thus we assume that taking some care is always worth the saved expected accident costs. Because of the discontinuity introduced by the negligence standard, we need a stronger assumption for the upper bound, that

$$\lim_{x \to 1} A(x,y) = -\infty \qquad \text{for all } y, \tag{2}$$

This would follow if it were financially impossible to achieve a care level of one. Given these assumptions, we can describe the supply of care in activity one as a function of the care taken in activity two by the first-order condition for the unconstrained maximization of A,

$$A_x(x,y) = 0. \tag{3}$$

Let us denote the solution to this equation by

$$x = x_1(y). \tag{4}$$

To give an explicit structure to the problem and to limit the proliferation of cases which would need analysis, we shall assume that increased care in activity two decreases the need to take care in activity one. That is, we assume that

$$A_{xy}(x,y) < 0 \qquad \text{for all } x,y. \tag{5}$$

The supply of care is thus decreasing in the care in activity two, as can be seen by implicit differentiation of the first-order condition, (3),

$$x'_1(y) = -\frac{A_{xy}}{A_{xx}} < 0, \tag{6}$$

where x'_1 is the derivative of x_1 with respect to y. In addition we shall assume that more care in activity two increases expected utility in activity one

$$A_y > 0. \tag{7}$$

Clearly, only some types of care decrease expected costs for both parties to potential accidents. At first glance, (5) and (7) may seem contradictory. But it is natural to have increased care by others decrease expected accident costs, and so increase expected utility, and yet to have the expected marginal utility of care decreased by increased care on the part of others, precisely because there is less chance of an accident, and so, less need for care.

We shall make a similar set of assumptions for activity two. Parallel to A is $D(x,y)$, which gives expected utility net of accident costs for someone in activity two. For later use we write D as $B(x,y) - C(x,y)$. This division has no role in the present section. Below, C will represent expected recognizable costs. Thus $B(x,y)$ is the expected utility ignoring accident costs which might be compensated (i.e., recognized costs), of someone in activity two who takes care level y when everyone in activity one is taking care level x. We assume that B is strictly concave in y, is twice continuously differentiable in both variables, and has

$$B_x(x,y) > 0, \qquad B_{xy}(x,y) < 0 \tag{8}$$

$$B_y(x,0) > 0 \quad \text{and} \quad \lim_{y \to 1} B(x,y) = -\infty. \tag{9}$$

Thus, $C(x,y)$ is the expected value of accident costs to someone in activity two which the legal system recognizes and for which compensation might be paid by those in activity one to those in activity two. At present this compensation is not paid, so utility is $B - C$. We assume that C is twice continuously differentiable in both variables and convex in each variable as well as satisfying

$$C > 0, C_x < 0, C_y < 0, C_{xy} > 0. \tag{10}$$

In the absence of any system transferring accident costs between parties to an accident, those in activity two each maximize $B(x,y) - C(x,y)$. This gives the first-order condition

$$B_y(x,y) - C_y(x,y) = 0. \tag{11}$$

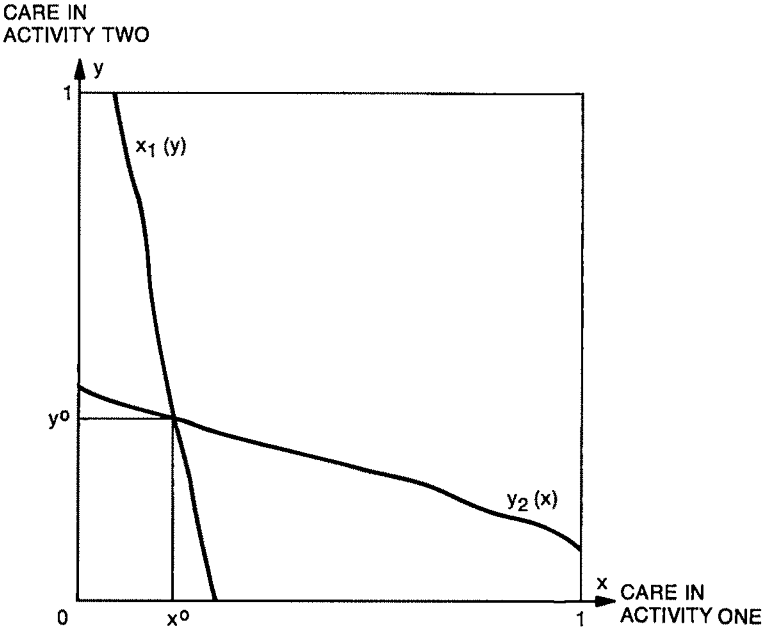
This gives the maximizing level of y as a function of x. Let us call it $y_2(x)$. Again, we have assumed that care decreases in the care taken in the other activity

$$y'_2(x) = -\frac{B_{xy} - C_{xy}}{B_{yy} - C_{yy}} < 0. \tag{12}$$

For equilibrium we have a simultaneous solution of the two first-order conditions (3) and (11). To see the structure (and to explain further assumptions) let us plot the two supply curves of care (reaction functions) in a single diagram. We shall denote the equilibrium care values by x^0 and y^0 and show them in Figure 1.

FIGURE 1

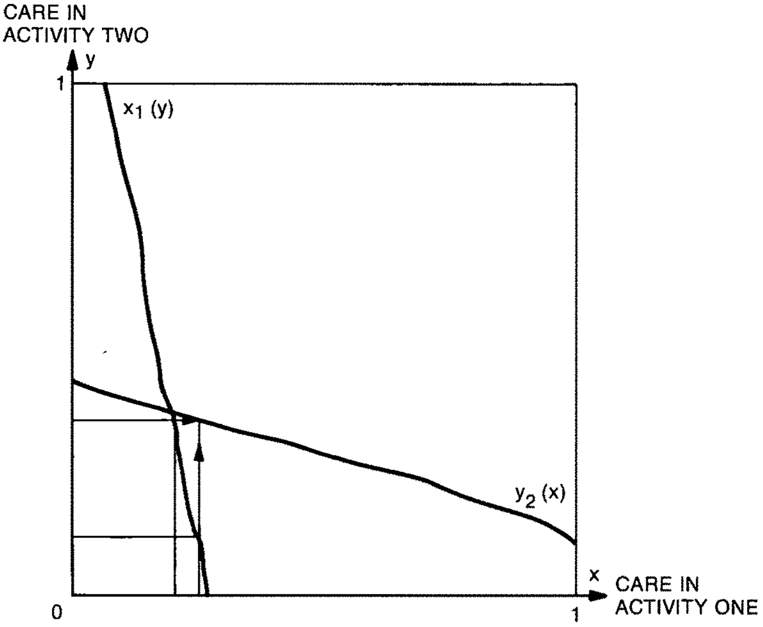
NO-LIABILITY EQUILIBRIUM



Given that the care of each is decreasing in the care of the other, there is the possibility of multiple equilibria. However, we assume the existence of a unique equilibrium. With reaction functions defined for all care levels in the other activity, the equilibrium is necessarily stable under the adjustment process of sequential choice in the two activities, each choice presuming constancy in the care taken in the other activity. Such an adjustment is shown in Figure 2.

FIGURE 2

ADJUSTMENT TO NO-LIABILITY EQUILIBRIUM



4. Individual choice

Choice with a negligence system. Now let us reexamine individual choice when there is the possibility that, after an accident, the person engaged in activity one must pay some of the costs of the person who was engaged in activity two at the time of the accident. If at the time of the accident the person in activity one was taking a level of care below the due care standard, $x < d_x$, he will be judged negligent. If, in addition, the person in activity two was taking a care level at least as great as the due care standard $y \geqq d_y$ (i.e., he is not contributorily negligent), then the person in activity one must pay those accident costs of the person in activity two which are recognized by the legal system. In a uniform equilibrium, all accidents will result in the same legal situation. Thus, in expected value terms, the expected costs $C(x,y)$ get shifted from the typical person in activity two to the typical person in activity one when the situation occurs which was described above.

Let us denote by U the expected utility net of all expected accident costs borne by a person in activity one. Then we have

$$U = \begin{cases} A(x,y) & : x \geqq d_x \quad \text{or} \quad y < d_y \\ A(x,y) - C(x,y) & : x < d_x \quad \text{and} \quad y \geqq d_y. \end{cases} \qquad (13)$$

We have denoted by $x_1(y)$ the care level which maximizes $A(x,y)$, ignoring the question of which parts of A are relevant for the utility function U. Let us denote by $x_2(y)$ the level of care which maximizes $A(x,y) - C(x,y)$. That is $x_2(y)$ satisfies

$$A_x(x,y) - C_x(x,y) = 0. \qquad (14)$$

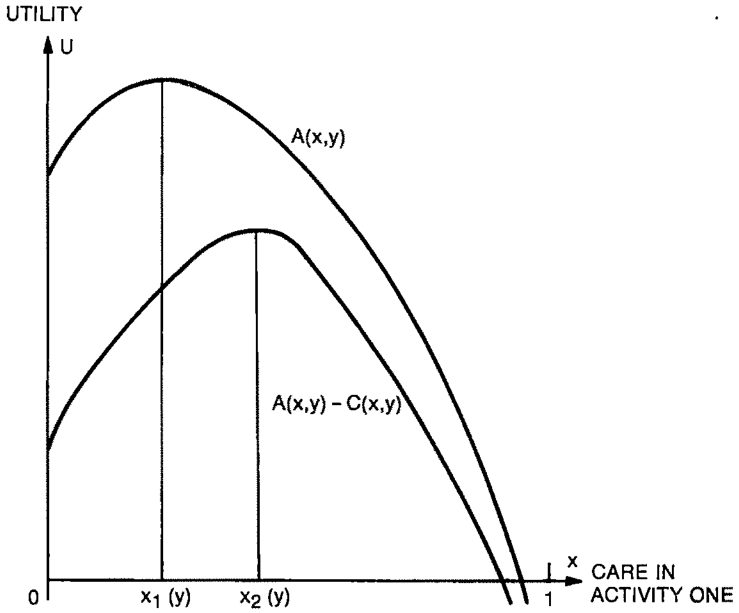
From the assumptions above x_2 decreases with y. In addition, given the concavity of A and convexity of C in x, x_2 exceeds x_1

$$x_2(y) > x_1(y). \tag{15}$$

That is, an individual who also bears the costs $C(x,y)$ takes more care than one who just bears the costs implicit in the definition of $A(x,y)$. Given the assumptions on the structure of the functions, they bear the relationship shown in Figure 3. To determine the relevant portions of the two functions we must evaluate contributory negligence (compare y and d_y) and locate the due care standard d_x on the diagram. For $y \geqq d_y$, U coincides with $A - C$ below d_x and with A above it. For $y < d_y$, U coincides with A and leads to the choice of $x_1(y)$. To consider the different cases with $y \geqq d_y$, let us consider Figure 3 with d_x in different positions. For $d_x < x_1(y)$, U is the dotted curve in Figure 4 and the maximizing level of care is clearly $x_1(y)$. If d_x lies between x_1 and x_2, the maximizing level of care is precisely the due care standard as in Figure 5. When the due care standard exceeds x_2, it is necessary (for determining choice) to compare the due care standard with the level of care which gives the same utility value $A(x,y)$ as the maximum level of $A(x,y) - C(x,y)$. Let us denote this level of care by $\hat{d}_x(y)$. It is defined by

FIGURE 3

EXPECTED UTILITY



$$A(\hat{d}_x,y) = A(x_2(y),y) - C(x_2(y),y), \tag{16}$$

where x_2 maximizes $A - C$ for the given y. The determination of $\hat{d}_x$ is shown in Figure 6. From the figure we can see that a due care level between x_2 and $\hat{d}_x$ leads to a choice of precisely the due care standard. A due care standard in excess of $\hat{d}_x$ leads the individual to choose to be negligent and to select the optimizing care level $x_2(y)$.

FIGURE 4

EXPECTED UTILITY

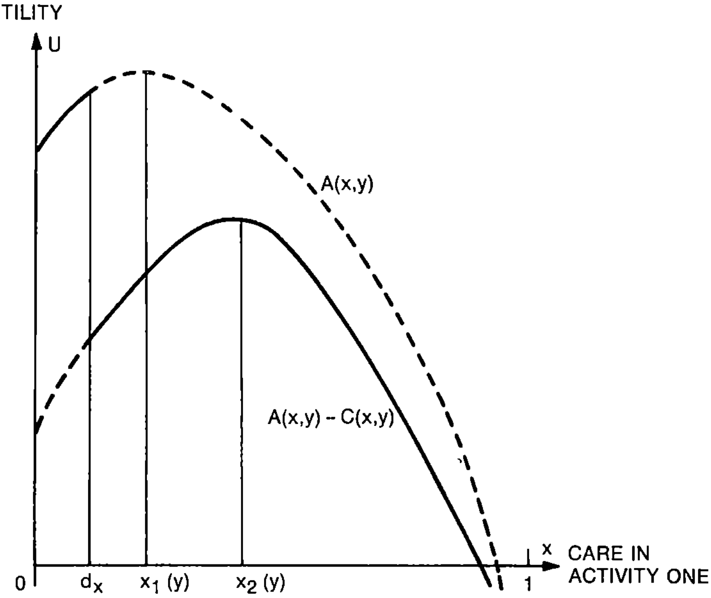
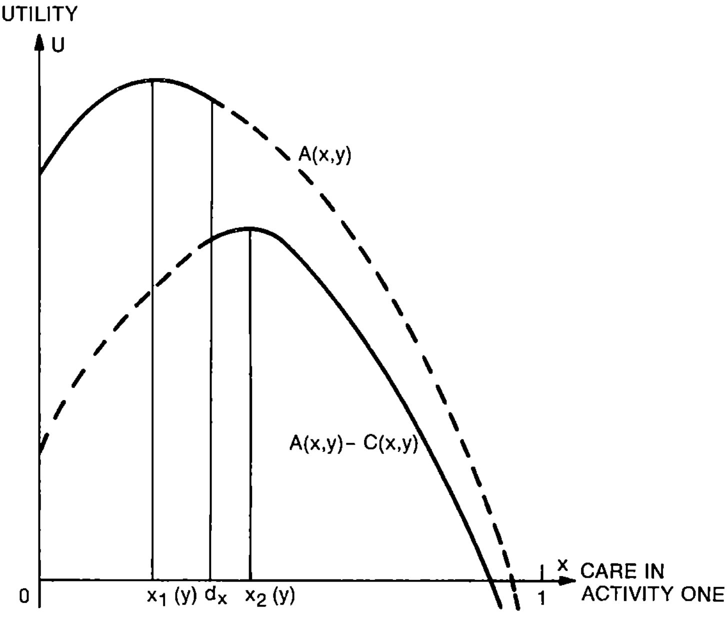


FIGURE 5

EXPECTED UTILITY



We can summarize the cases relating choice to the due care standard and the situation in the other activity in Table 1.

In the absence of a legal system, choice was $x_1(y)$. Thus with the legal system, care in activity one is at least as large as it would be without the legal system for a given level of care in activity two. If $y \geqq d_y$ and $d_x > x_1(y)$, then care in activity one is strictly greater than it would be without a legal system.

For later analysis, let us explore the behavior of $\hat{d}_x(y)$ as a

FIGURE 6

EXPECTED UTILITY

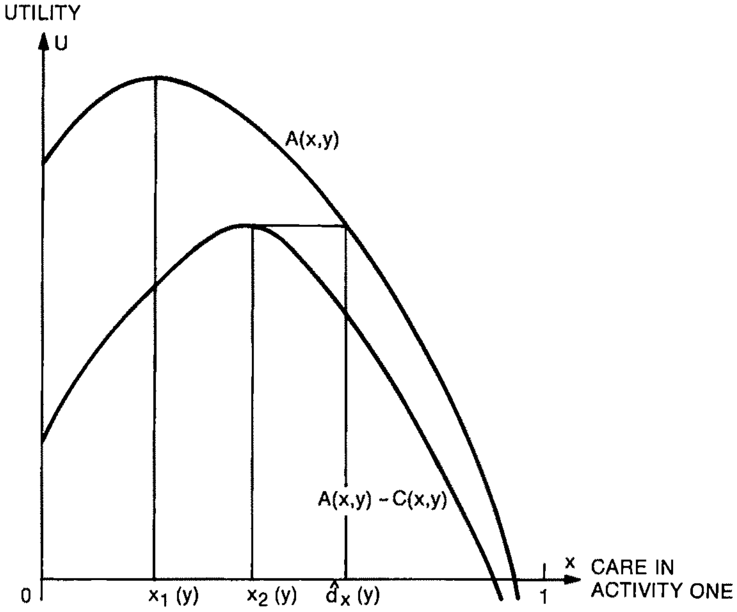


TABLE 1

CHOICE IN ACTIVITY ONE

DUE CARE STANDARD	ANY	$d_x \leqq x_1$	$x_1 \leqq d_x \leqq \hat{d}_x$,	$\hat{d}_x \leqq d_x$
CARE IN ACTIVITY TWO	$y < d_y$	$y \geqq d_y$	$y \geqq d_y$	$y \geqq d_y$
OPTIMAL LEVEL OF CARE IN ACTIVITY ONE	$x_1(y)$	$x_1(y)$	d_x	$x_2(y)$

function of y. Implicitly differentiating the definition of $\hat{d}_x$ (and using the first-order condition for x_2,

$$\hat{d}'_x(y) = -\frac{A_y(\hat{d}_x,y) - A_y(x_2,y) + C_y(x_2,y)}{A_x(\hat{d}_x,y)}. \qquad (17)$$

From Figure 6 we know that A_x is negative at $\hat{d}_x$ and that $\hat{d}_x$ exceeds x_2 [that is, we are not interested in the other solution to (16)]. From the assumptions above, $C_y < 0$ and $A_{xy} < 0$. Thus, since $\hat{d}_x > x_2$,

$$A_y(\hat{d}_x,y) < A_y(x_2,y). \qquad (18)$$

Thus $\hat{d}_x$ is decreasing in y.

☐ **Choice with contributory negligence.** In an exactly parallel fashion we can describe the choice of an individual engaged in activity two. If he is contributorily negligent, or if the person with whom he has an accident is not negligent, he must bear the accident costs which the legal system will otherwise shift to the other party. Denoting his expected utility by V, we have

$$V = \begin{cases} B(x,y) & : x < d_x \quad \text{and} \quad y \geqq d_y \\ B(x,y) - C(x,y) & : x \geqq d_x \quad \text{or} \quad y < d_y. \end{cases} \qquad (19)$$

We have defined $y_2(x)$ as the maximizing value for $B - C$. We define $y_1(x)$ as the maximizing value for $B(x,y)$. From the assumptions above we have

$$y'_1(x) = -\frac{B_{xy}}{B_{yy}} < 0, \ y_1(x) < y_2(x). \qquad (20)$$

We can define $\hat{d}_y(x)$ as the level of care equating the value of B to the maximizing value of $B - C$:

$$B(\hat{d}_y,x) = B(y_2(x),x) - C(y_2(x),x). \qquad (21)$$

By implicit differentiation and a parallel argument to that above we have $\hat{d}_y(x)$ decreasing in x. We could now repeat the diagrams of the previous section replacing A by B and x by y. Instead, let us merely summarize the optimal choice in Table 2.

TABLE 2

CHOICE IN ACTIVITY TWO

DUE CARE STANDARD	ANY	$d_y \leq y_1$	$y_1 \leq d_y \leq \hat{d}_y$	$\hat{d}_y \leq d_y$
CARE IN ACTIVITY ONE	$x \geq d_x$	$x < d_x$	$x < d_x$	$x < d_x$
OPTIMAL CARE IN ACTIVITY ONE	$y_2(x)$	$y_1(x)$	d_y	$y_2(x)$

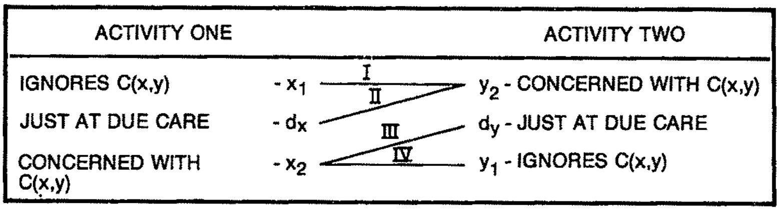
In contrast to the situation in activity one, the presence of the legal system may raise or lower the level of care chosen in activity two in response to a given amount of care in activity one. Only if $x < d_x$ and $y_2(x) < d_y \leqq \hat{d}_y$, is care in activity two increased.

5. Uniform equilibria

■ **Types of uniform equilibria.** For any pair of due care standards, (d_x,d_y), we need to repeat the analysis done for the no-liability equilibrium—determine the supply curves of care in the two activities, check to see whether they can be simultaneously solved (check whether there is an equilibrium with correct perceptions), and characterize the equilibrium. In Tables 1 and 2 above we have described the reaction functions ot the individuals in terms of the due care standards and the care in the other activity. There were three types of maximizing choices for each individual—precisely the due care standard (d_x or d_y), an internal maximum under the assumption that court recognized costs $C(x,y)$ remain on participants in activity two (x_1 or y_2), and an internal maximum assuming that the court recognized costs are shifted to activity one (x_2 or y_1). With three types of choices in each activity there are potentially nine types of equilibria. However, there are only four types of equilibria that are in fact possible, as marked in Table 3. The restriction on types of equilibria comes from the fact that the costs $C(x,y)$ are borne in one and only one activity. Thus it cannot be true that parties in neither activity or both activities are concerned with $C(x,y)$. Furthermore, an individual will only go to the expense of being precisely at the due care standard—ignoring the borderline where the standard coincides with the internal maximizing choice—to avoid bearing these transferable costs. Thus

TABLE 3

TYPES OF EQUILIBRIA



the costs must be borne in the other activity in this case. This gives us the four types of equilibria marked in Table 3. We shall begin by considering type I which coincides with the equilibrium allocation in the absence of a legal system. Then, since the analysis is similar, we shall consider type IV before turning to types II and III.

☐ **No-liability equilibrium under negligence.** For some pairs of due care standards, the resulting equilibrium is precisely the same as that which would occur in the absence of a liability system.[15] Without any possibility of shifting costs, those in activity one maximize $A(x,y)$ while those in activity two maximize $B(x,y) - C(x,y)$. This leads to the choices $x = x_1(y)$ and $y = y_2(x)$. The simultaneous solution of these two equations we have denoted by (x^0,y^0). The simultaneous choices of x_1 and y_2 correspond to an equilibrium of type I in Table 3. Now let us see what pairs of due care standards can lead to this type of equilibrium.

If the due care standard in activity one is below x^0, the choice of x^0 by everyone in activity one leaves no one negligent. If there is no negligence in activity one, an individual in activity two must bear his own accident costs whether contributorily negligent or not. Thus each individual will select $y_2(x^0) = y^0$ for any value of d_y. (See column one of Table 2.) Given that everyone in activity two has selected y^0, each person in activity one will select $x_1(y^0) = x^0$, whenever this choice does not imply negligence (columns one and two of Table 1). Thus we will have an equilibrium at (x^0,y^0) for any pair of due care standards (d_x,d_y) satisfying $d_x \leqq x^0$.

This description covers two ways of achieving this equilibrium—by having no one negligent and by having contributory negligence in activity two, but no negligence in activity one. There is a third way of achieving this equilibrium—by having the due care standard in activity two sufficiently high that, despite negligence in activity one, everyone in activity two prefers the best negligent care level to any nonnegligent care level. (This corresponds to column 4 of Table 2.) For this type of equilibrium we require $d_x > x^0$ so that everyone in activity one would be negligent if he chose x^0. For due care standards in activity two above $\hat{d}_y(x^0)$

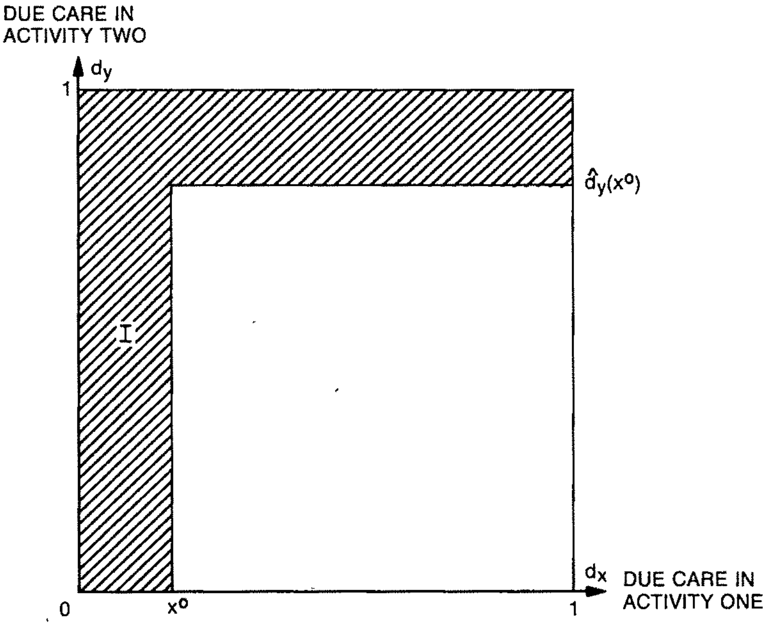
---

[15] In an administratively costless model, no liability (and strict liability) arise as equilibria with a negligence system and particular due care standards. One would expect a negligence system to be much more expensive to administer than either no or strict liability. A full efficiency analysis would consider administrative costs in addition to the efficiency questions analyzed in this paper.

everyone in activity two will select y^0 and thus, being negligent, justify the choice of x^0 by those in activity one. We depict the region of due care standards giving rise to equilibrium at (x^0,y^0) in Figure 7. Within the region, changes in the due care standards have no effects on equilibrium.

FIGURE 7

REGION OF EQUILIBRIUM OF TYPE ONE



☐ **Strict liability equilibrium.** If there were strict liability for those in activity one, they would bear the recognized accident costs $C(x,y)$ independent of the pattern of behavior of the two parties at the time of an accident. Thus, in activity one the utility function would be $A(x,y) - C(x,y)$, while in activity two the utility function would be $B(x,y)$. These utility functions would result in optimizing care choices

$$\begin{aligned} x &= x_2(y) \\ y &= y_1(x). \end{aligned} \tag{22}$$

Let us denote the simultaneous solution of these equations by (x^{00},y^{00}). As with the absence of liability, we assume a unique equilibrium and have the implied stability condition.

We can compare this equilibrium to that without liability by considering the reaction functions for both cost allocations in a single diagram in Figure 8. As is clear from the diagram we have

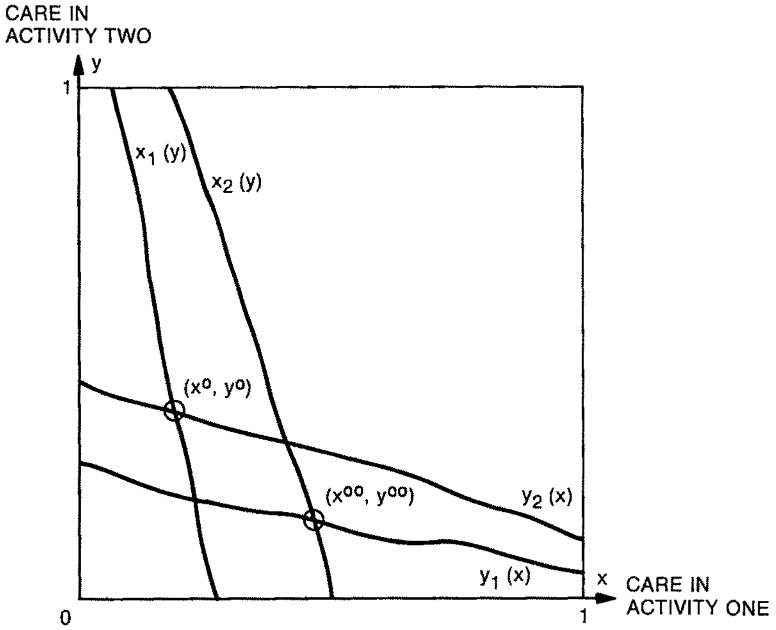
$$\begin{aligned} x^0 &< x^{00} \\ y^0 &> y^{00}. \end{aligned} \tag{23}$$

Thus the shifting of costs from activity two to activity one increases the equilibrium level of care in the latter while decreasing it in the former.

Now let us consider the values of the due care standards which will result in this equilibrium. With a negligence-contributory neg-

FIGURE 8

REACTION FUNCTIONS



ligence system the shifting of costs between activities requires that those in activity two satisfy the due care standard, i.e.,

$$d_y \leqq y^{00}. \tag{24}$$

Given this condition and the choice of y^{00} in activity two, those in activity one will bear the shiftable costs when they prefer the best negligent care level to any nonnegligent care level. This can occur for sufficiently high due care standards, i.e., those satisfying

$$d_x \geqq \hat{d}_x(y^{00}). \tag{25}$$

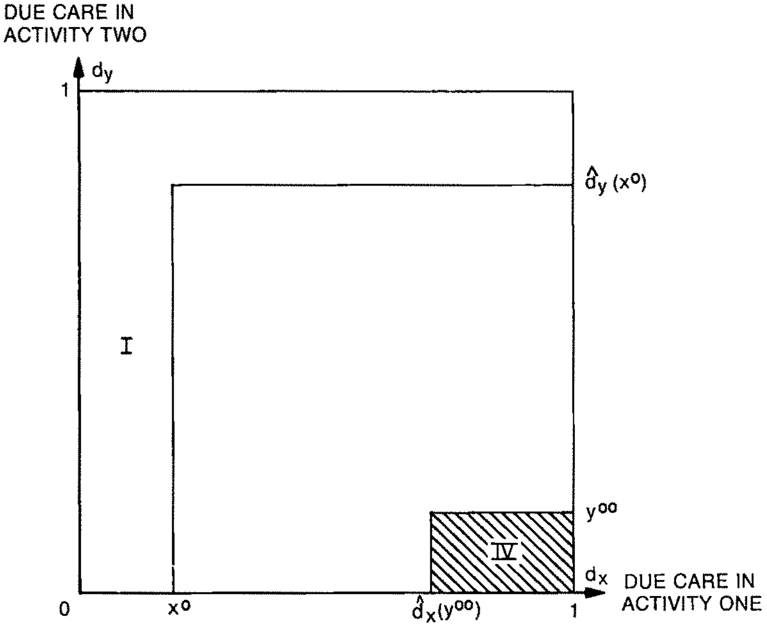
(See column 4 of Table 1.) For such a due care standard those in activity one select $x^{00} = x_2(y^{00})$, justifying the choice of y^{00} in activity two. We depict this region of equilibria, along with that of type one in Figure 9. As was true in region I, equilibrium is independent of the values of the due care standards, provided we remain within region IV.

☐ **Due care in activity one.** For some combinations of due care standards, individuals in activity one will select a care level precisely equal to the due care standard (even when it does not coincide with x_1). For them to have an incentive to take precisely this care level to avoid negligence, they must be avoiding the court imposed costs. Thus, it must be true that those in activity two are not contributorily negligent. Because those in activity one are choosing to be nonnegligent, those in activity two must bear these costs, $C(x,y)$, and must be selecting a care level reflecting the full costs that they must carry; i.e., they must be selecting $y_2(d_x)$ when those in activity one choose d_x. Thus one constraint that must be satisfied for equilibrium of type II is that

$$y_2(d_x) \geqq d_y. \tag{26}$$

FIGURE 9

REGION OF TYPE IV EQUILIBRIUM



From Table 1 we see that those in activity one will select the due care standard if it exceeds the care they would take ignoring the costs $C(x,y)$ and if it is not so high that a higher utility can be achieved by being negligent. Thus, we have the two constraints

$$x_1[y_2(d_x)] \leqq d_x \leqq \hat{d}_x[y_2(d_x)]. \qquad (27)$$

Let us examine the constraints on d_x in detail. We know that the lower bound inequality is satisfied for d_x equal to x^0, since

$$x^0 = x_1[y_2(x^0)]. \qquad (28)$$

Since we have assumed a unique equilibrium in the absence of a liability system, this will be the only value of d_x satisfying this equation. From Figure 1 we can see that the inequality is satisfied for any larger value of d_x. Thus the lower bound inequality in (27) becomes

$$x^0 \leqq d_x. \qquad (29)$$

Without further assumptions we cannot rule out the possibility of several solutions to the equation

$$d_x = \hat{d}_x[y_2(d_x)]. \qquad (30)$$

This could give a complicated shape to this region of equilibria. To avoid this complication we assume a single solution which we shall denote $\bar{d}_x$. Thus, the inequality in (27) is satisfied for all lower values of d_x. Combining these results we see that region II is defined by three inequalities

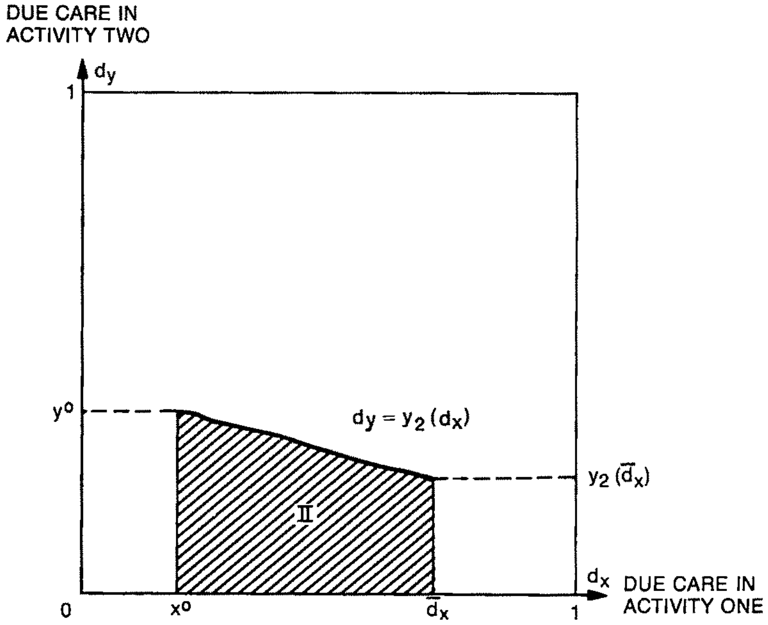
$$x^0 \leqq d_x \leqq \bar{d}_x,\ d_y \leqq y_2(d_x), \qquad (31)$$

and can be depicted as in Figure 10.

Relating the position of region II to that of the other regions, we see that it starts on the edge of region I. However, without

FIGURE 10

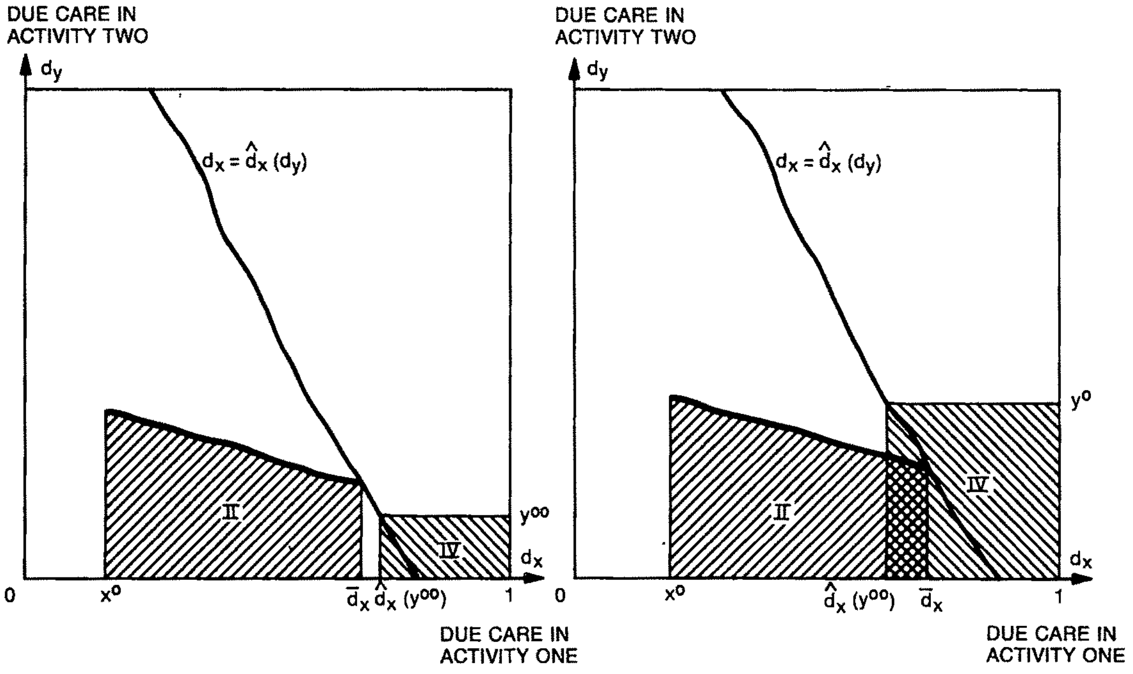
REGION OF TYPE II EQUILIBRIUM



further assumptions, there is no necessary relationship between d_x and $\hat{d}_x(y^{00})$, the lower bound constraint for region IV. However, the points $[\bar{d}_x, y_2(\bar{d}_x)]$ and $[\hat{d}_x(y^{00}), y^{00}]$ both lie on the curve $x = \hat{d}_x(y)$ which cuts $y = y_2(x)$ from above by the assumptions we have just made. Thus, one of the points is northwest of the other, allowing only the two possibilities shown in Figure 11. Therefore, there may be a gap between the two regions or an overlap allowing two different equilibria for some pairs of due care

FIGURE 11

EQUILIBRIA OF TYPES II AND IV

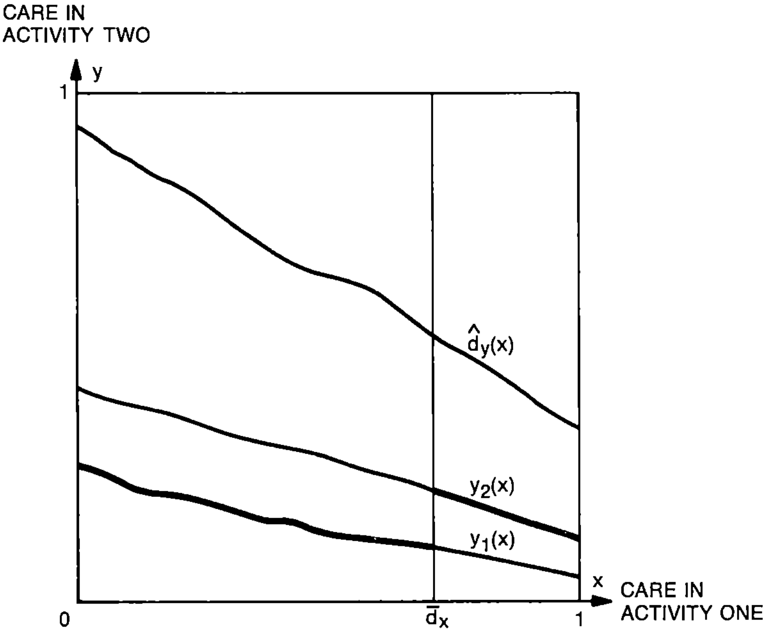


standards. To see this situation more clearly, let us examine the reaction curves for a point in region II. In particular, let us consider the reaction curves for the values of the due care standards $(\bar{d}_x,0)$. (We do not examine the point at the upper end, $[\bar{d}_x,y_2(\bar{d}_x)]$, since that point will be in region III if it is not in region IV.)

Let us start with the reaction curve in activity two. We seek to describe the chosen y as a function of x. For x below d_x, those in activity two will select y_1 provided that y_1 exceeds d_y as it necessarily will here. (See Table 2.) Once x equals or exceeds d_x, those in activity two must bear the costs $C(x,y)$ and so will choose $y_2(x)$. Thus, the reaction curve appears as the heavy line in Figure 12.

FIGURE 12

REACTION CURVE IN ACTIVITY TWO

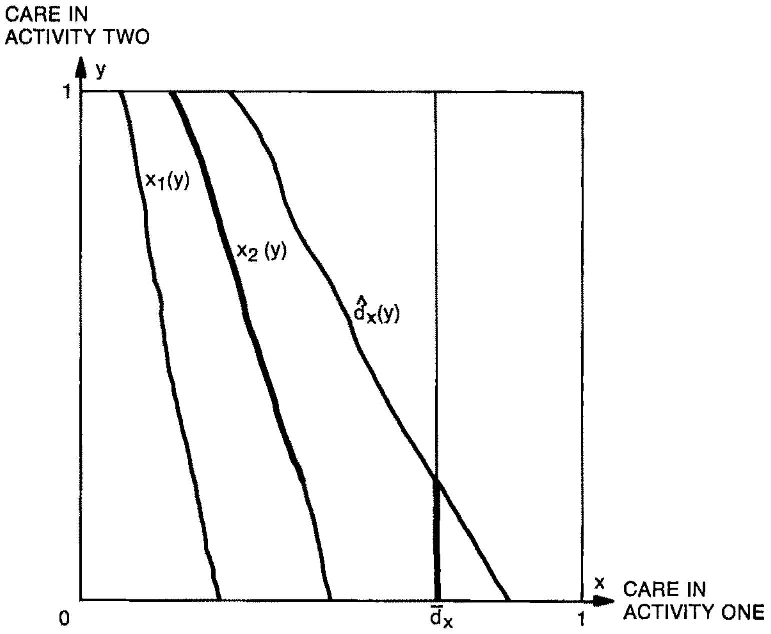


Similarly, if there were values of y below d_y, those in activity one would not bear the costs $C(x,y)$ and thus would select $x_1(y)$. For y equal to d_y, since $\bar{d}_x < \hat{d}_x(0)$ everyone selects d_x. For higher values of y, they continue to select d_x until d_x exceeds $\hat{d}_x(y)$. Then x_2 is chosen. Thus, the reaction curve is the heavy line in Figure 13.

Looking to combine Figures 12 and 13, we note that $\bar{d}_x$ satisfies $\bar{d}_x = \hat{d}_x[y_2(\bar{d}_x)]$. Thus there is an equilibrium with $x = d_x$ and $y = y_2(d_x)$. There will be a second equilibrium for these due care standards when $x_2[y_2(d_x)]$ is less than the value of x which equates $y_1(x)$ with $y_2(\bar{d}_x)$. This is shown in Figure 14. The key element is the comparison of the horizontal differences between y_2 and y_1 and between $\hat{d}_x$ and x_2. In the other case, where the horizontal distance between $\hat{d}_x$ and x_2 is less than that between y_2 and y_1, we will have no uniform equilibrium for a value of d_x slightly larger than $\bar{d}_x$ and a value of d_y of zero. An example is shown in Figure 15. Summarizing the analysis of equilibria of type II, we see that x increases precisely in step with d_x within the region. The equi-

FIGURE 13

REACTION CURVE IN ACTIVITY ONE



librium value of y decreases with d_x. Changes in d_y have no effects. Recognized accident costs are borne by those in activity two.

□ **Due care in activity two.** In a parallel fashion we have equilibrium of type III when those in activity one are negligent and those in activity two are expending the effort to be precisely at the due care level. With those in activity two at d_y, those in activity one will choose to be negligent when the due care standard

FIGURE 14

EXAMPLE OF MULTIPLE EQUILIBRIA

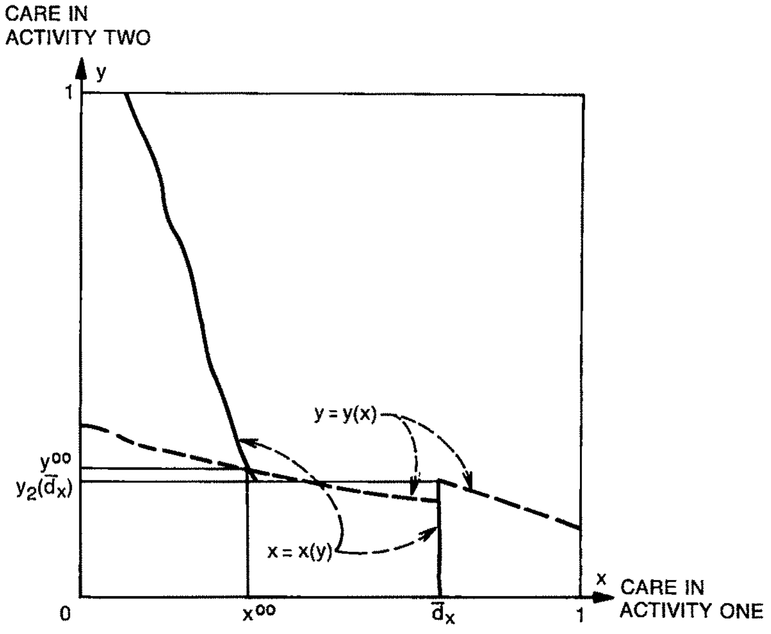
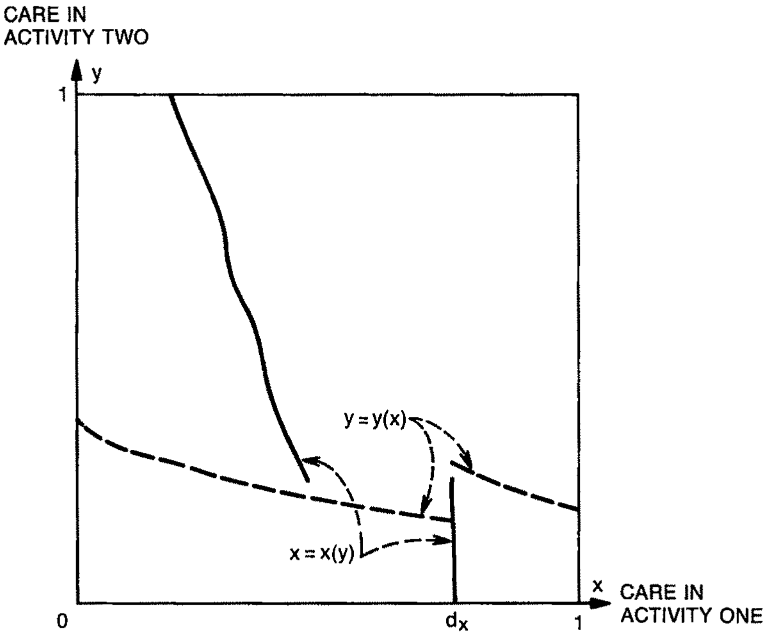


FIGURE 15

EXAMPLE OF NO UNIFORM EQUILIBRIA



is sufficiently high; i.e., when it satisfies

$$d_x \geqq \hat{d}_x(d_y). \tag{32}$$

In this case, those in activity one will select a care level $x_2(d_y)$. From Table 2 we know that those in activity two will select the due care level [given that those in activity one are at $x_2(d_y)$], provided it satisfies the constraints

$$y_1[x_2(d_y)] \leqq d_y \leqq \hat{d}_y[x_2(d_y)]. \tag{33}$$

Let us begin with the lower bound constraint. For d_y equal to y^{00} we know that $y_1[x_2(d_y)]$ is equal to d_y. From the uniqueness of equilibrium with shifted costs we know that the inequality is satisfied for any greater value of d_y and unsatisfied for any lower value. Since the upper bound of region IV was also defined by y^{00}, regions III and IV share a common boundary. From Figure 11 we see that there is no overlap between regions II and III.

Without further assumptions we cannot rule out the possibility of several solutions, d_y, which just satisfy the upper bound constraint. To avoid complicating the possible patterns of equilibrium even further, we assume the existence of a unique value, $\bar{d}_y$, satisfying

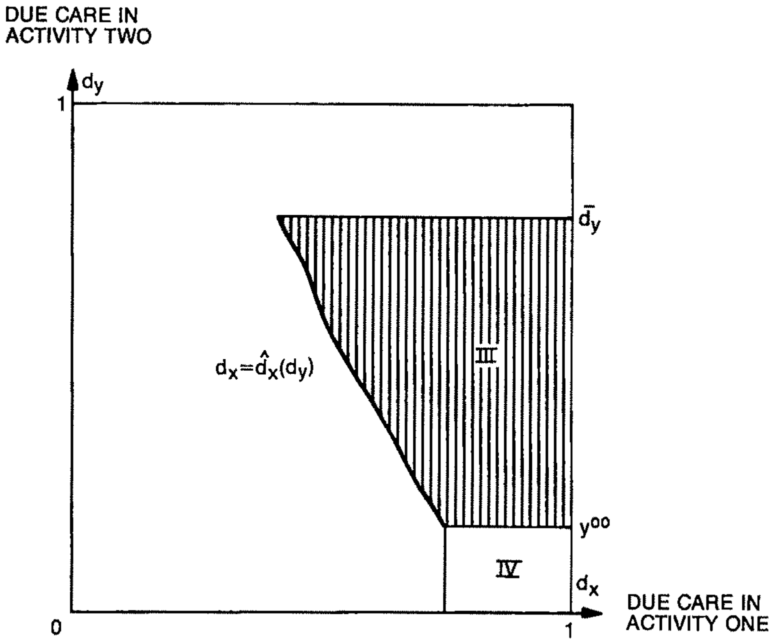
$$\bar{d}_y = \hat{d}_y[x_2(\bar{d}_y)]. \tag{34}$$

Thus, the inequality in (33) is satisfied for lower values of d_y and unsatisfied for greater ones. Therefore, region III has the shape shown in Figure 16.

It remains to relate the position of region III to that of region I. Without further assumptions there are three possible configurations. These are shown in Figure 17. To see that these are the different possible configurations, let us consider diagrams showing

FIGURE 16

REGION OF TYPE III EQUILIBRIUM



the different relative positions of the relevant curves. Three possible configurations are shown in Figures 18 to 20, depending on the relative vertical distance between $\hat{d}_y$ and y_2 compared with those between x_1 and x_2 and between x_1 and $\hat{d}_x$. To compare the boundaries we need to compare the values of d_x and d_y which are at the corners of the two regions —[x^0,$\hat{d}_y(x^0)$] and [$\hat{d}_x(\bar{d}_y)$,$\bar{d}_y$]. Figure 18 depicts the case of no overlap. Here, $\bar{d}_y$, defined by the intersection of $x_2(y)$ and $\hat{d}_y(x)$, is below $\hat{d}_y(x^0)$. It necessarily follows that x^0 is less than $\hat{d}_x(\bar{d}_y)$. In Figure 19, $x_1(y)$ and $x_2(y)$ are close to each other relative to $y_2(x)$ and $\hat{d}_y(x)$. This results in $\bar{d}_y$ exceeding $\hat{d}_y(x^0)$. Since $\hat{d}_x(y)$ is above $x_1(y)$ by more than $\hat{d}_y$ is above y_2, we still have x^0 less than $\hat{d}_x(\bar{d}_y)$. In Figure 20 we depict the remaining case where $\bar{d}_y$ exceeds $\hat{d}_y(x^0)$ and x^0 exceeds $\hat{d}_x(\bar{d}_y)$. In a parallel fashion to the analysis above, we could consider reaction curves to show the different possible numbers of equilibria. Summing up, in region III accident costs are borne in activity one; changes in d_y affect equilibrium but changes in d_x do not.

6. Efficient bearing of recognized costs

■ Considering the possible patterns of uniform equilibria, there are six possibilities (combining Figures 11 and 17). Arbitrarily selecting one of them (that which has gaps), we show the choice of care occurring in the different regions in Figure 21. Having this picture of possible equilibria choices will serve as a background for the discussion of efficiency. Let us also summarize the situation in Table 4.

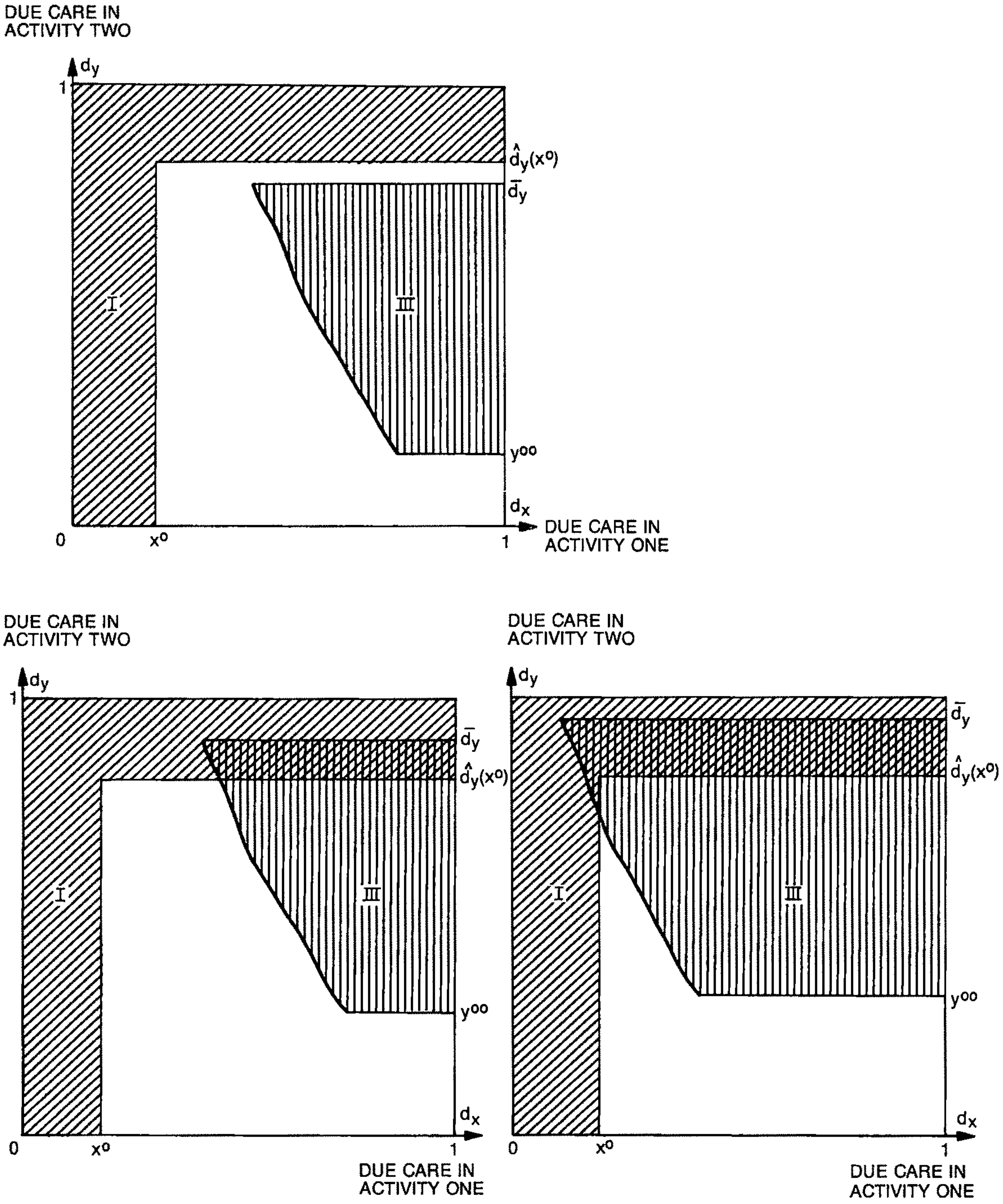
Now let us consider two approaches to efficiency relative to the possible uniform equilibrium in the economy. For full efficiency[16]

---

[16] Since each utility function is linear in income (being linear in acci-

FIGURE 17

RELATIONSHIP OF EQUILIBRIA OF TYPES I AND III



in this economy of two activities we would want each person to make his care decisions in a way that reflects all the impacts of his care on accident costs, i.e., on $A + B - C$. Thus, the first-order conditions for full efficiency would be

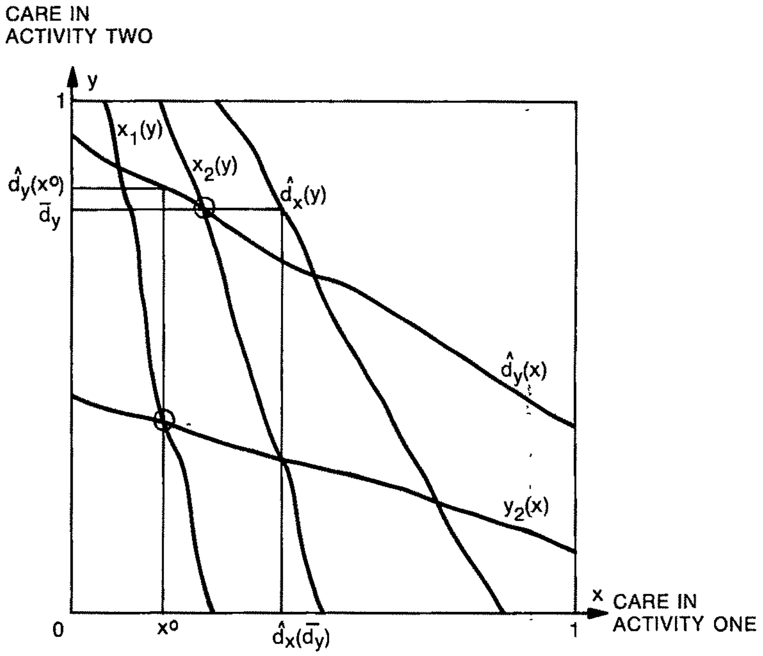
$$\begin{aligned} A_x + B_x - C_x &= 0 \\ A_y + B_y - C_y &= 0. \end{aligned} \qquad (35)$$

---

dent costs), any Pareto optimum which did not involve zero income for some party satisfies the same efficiency conditions.

FIGURE 18

EQUILIBRIA OF TYPES I AND III



However the legal system only recognizes the costs, $C(x,y)$, as possibly transferable between parties in the different activities. The externalities arising from A_y and B_x being nonzero at equilibrium are outside the range of issues which the legal system is considering. We shall define efficiency in the bearing of recognized costs as occurring when each party makes his care decision dependent on his own costs and on the recognized costs. Thus the equa-

FIGURE 19

EQUILIBRIA OF TYPES I AND III

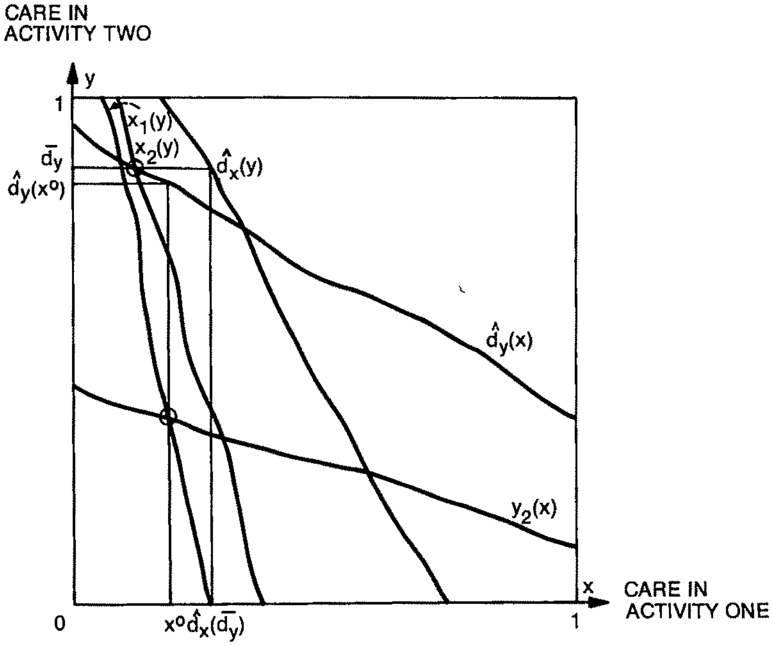
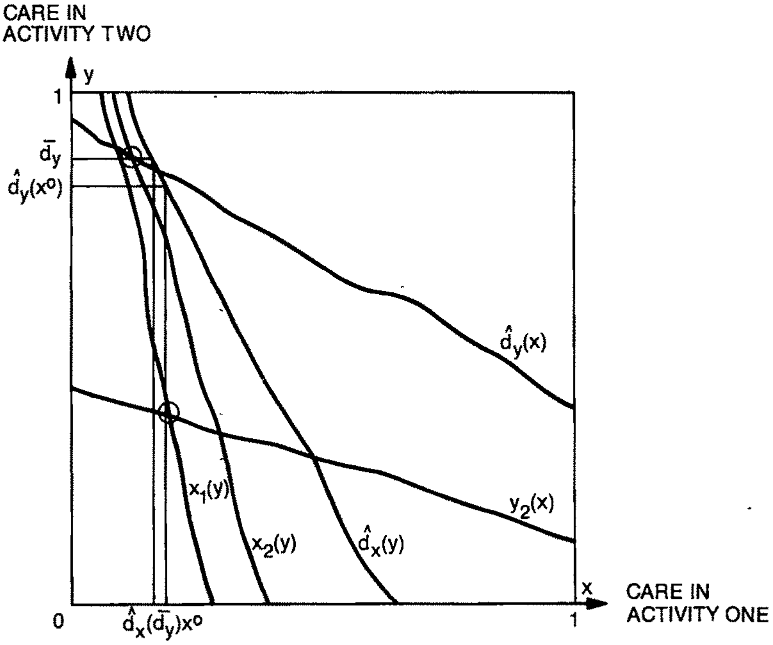


FIGURE 20

EQUILIBRIA OF TYPES I AND III



tions describing the efficient position in this case are

$$\begin{aligned} A_x - C_x &= 0 \\ B_y - C_y &= 0. \end{aligned} \tag{36}$$

Let us denote by (x^*, y^*) the solution to these two equations. In this position each party is evaluating his marginal impact on the recognized costs $C(x,y)$ and each party is examining his own costs which are not recognized. Obviously, if there are no unrecognized externalities, the two concepts coincide.

FIGURE 21

CHOICE OF CARE IN REGIONS I-IV

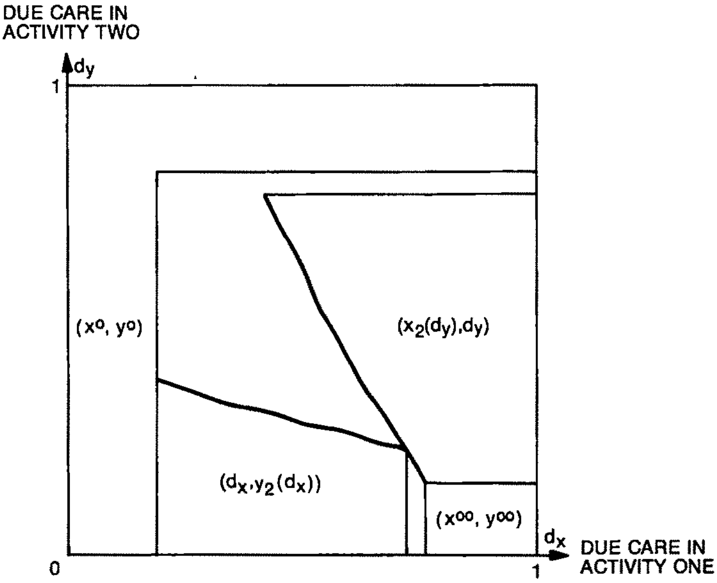


TABLE 4

POSSIBLE EQUILIBRIA AND CHOICES OF CARE

EQUILIBRIUM TYPE	ACTIVITY BEARING ACCIDENT COSTS	CARE IN ACTIVITY ONE	CARE IN ACTIVITY TWO	DUE CARE STANDARDS IN THIS REGION
I	TWO	x^o	y^o	$d_x \leqq x^o$ OR $d_y \geqq \hat{d}_y(x^o)$
II	TWO	d_x	$y_2(d_x)$	$x^o \leqq d_x \leqq \bar{d}_x$ AND $d_y \leqq y_2(d_x)$
III	ONE	$x_2(d_y)$	d_y	$d_x \geqq \hat{d}_x(d_y)$ AND $y^{oo} \leqq d_y \leqq \bar{d}_y$
IV	ONE	x^{oo}	y^{oo}	$d_x \geqq \hat{d}_x(y^{oo})$ AND $d_y \leqq y^{oo}$

There are three efficiency propositions which can be made, given these definitions. First, efficiency in the bearing of recognized costs can be achieved by setting $d_x = x^*$ or $d_y = y^*$ provided that the other due care standard is set appropriately. (Setting both $d_x = x^*$ and $d_y = y^*$ will suffice.) Second, it is not possible to achieve an equilibrium at the full efficiency point when A_y and B_x are both positive for all values of x and y. Third, it is not generally true that the equilibrium at (x^*,y^*) is the most efficient equilibrium which can be attained in this economy; i.e., (x^*,y^*) does not generally maximize $A + B - C$ among the set of equilibria.

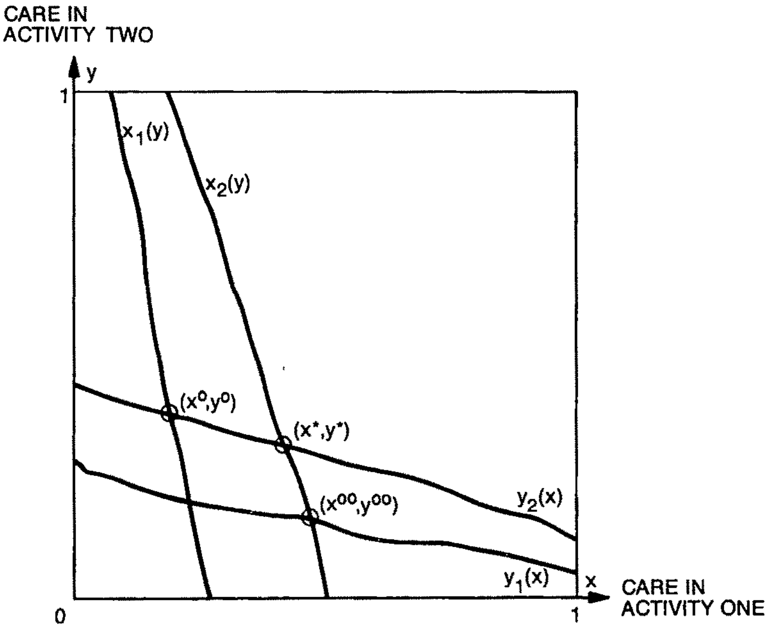
Let us begin by relating the efficient point to the equilibria with no-liability and strict liability. From the definitions of the optimizing care levels for different cost allocations it is clear that the efficient point satisfies the equations

$$\begin{aligned} x^* &= x_2(y^*) \\ y^* &= y_2(x^*). \end{aligned} \tag{37}$$

Thus we can depict the three positions in Figure 22 with the four

FIGURE 22

REACTION CURVES AND EFFICIENCY IN BEARING RECOGNIZED COSTS



choice functions. As is clear from the diagram (given the negative slopes of all four functions), the care levels satisfy

$$\begin{aligned} x^0 &< x^* < x^{00} \\ y^0 &> y^* > y^{00}. \end{aligned} \tag{38}$$

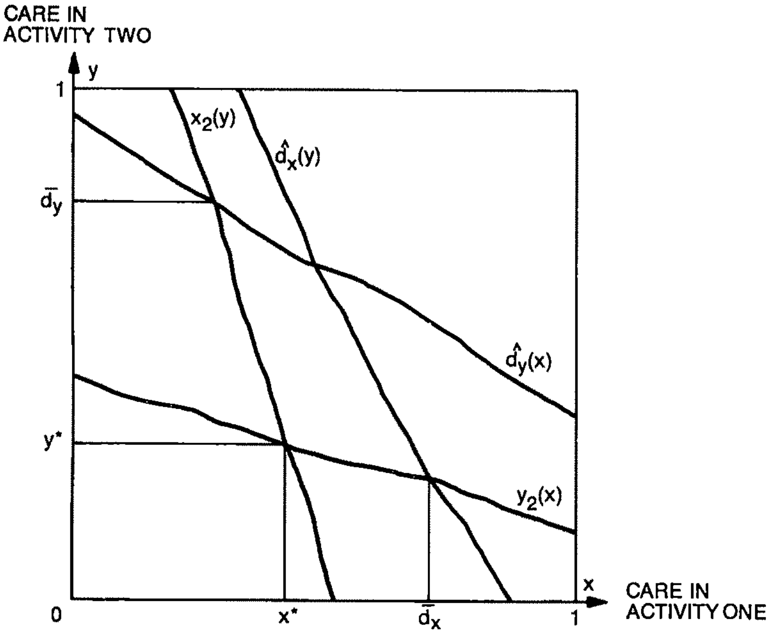
Thus efficiency is not attainable in regions I or IV.

In region II, the equilibrium care levels are $[d_x, y_2(d_x)]$. Thus, if x^* lies between x^0 and $\bar{d}_x$, it is possible to have the efficient equilibrium in region II by setting d_x equal to x^* and d_y less than or equal to $y_2(d_x)$ which equals y^*. Similarly, the efficient equilibrium can be achieved by a pair of due care standards in region III provided that y^* lies between y^{00} and $\bar{d}_y$. In this case we set d_y equal to y^* and d_x greater than $\hat{d}_x(y^*)$. From the inequalities on the equilibrium levels (38), we need only check that $x^* \leqq \bar{d}_x$ and $y^* \leqq \bar{d}_y$.

In Figure 23 we have drawn $x_2(y)$, $\hat{d}_x(y)$, and $y_2(x)$. Since

FIGURE 23

COMPARISON OF x^* AND $\bar{d}_x$

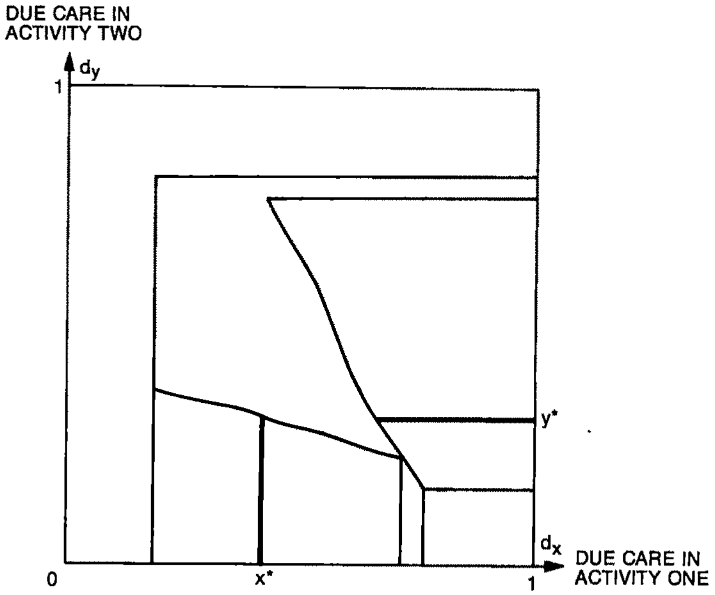


x^* occurs at the intersection of the first and third while $\bar{d}_x$ is at the intersection of the second and third [see (30)], we see that $x^* < \bar{d}_x$. Similarly, since y^* is at the intersection of $y_2(x)$ and $x_2(y)$ and $\bar{d}_y$ at that of $x_2(y)$ and $\hat{d}_y(x)$, we have the desired inequality. Thus it is always possible to achieve efficiency in the bearing of recognized costs and thus to achieve full efficiency if all costs are recognized. (We can also rule out the possibility of multiple uniform equilibria at these due care standards.) We now add these two loci to the diagram showing the regions in Figure 24.

To see that it is not possible to achieve full efficiency we note that all equilibrium care levels lie either on the curve $x = x_2(y)$ or on the curve $y = y_2(x)$. The full efficiency point lies on neither curve so long as A_y and B_x are not zero. Where these derivatives

FIGURE 24

EFFICIENT CARE LEVELS IN BEARING RECOGNIZED COSTS



are both positive, the point of full efficiency lies above and to the right of these two curves.

Within the interior of region II we can describe the sum of utilities as a function of d_x:

$$S(d_x) = A[d_x,y_2(d_x)] + B[d_x,y_2(d_x)] - C[d_x,y_2(d_x)]. \quad (39)$$

Calculating the derivative with respect to d_x, we have

$$S'(d_x) = A_x + B_x - C_x + (A_y + B_y - C_y)y'_2. \quad (40)$$

Evaluating this at x^*, we have

$$S'(x^*) = B_x + A_y y'_2. \quad (41)$$

Thus the point of efficient bearing of recognized costs will not have the highest sum of utilities in region II, if it is not true that

$$y'_2 = -\frac{B_x}{A_y}; \quad (42)$$

There is no reason in general why y_2', which satisfies (12), should also satisfy this equation at the point x^*.[17]

7. Nonuniform equilibria

■ We shall distinguish three types of nonuniform equilibria. In equilibrium of type V everyone in activity two chooses the same care level while those in activity one are divided between two different care levels. In type VI equilibrium we have the reverse case —uniformity in activity one and nonuniformity in activity two. In equilibrium of type VII we have nonuniformity in both activities simultaneously.

---

[17] I have not examined the question, asked by R. Whitcomb, of whether nonuniform equilibria have potentially greater sums of utilities than the best uniform equilibrium.

☐ **Nonuniformity in activity one.** Continuing the assumption that everyone in activity two selects the same care level, let us look for equilibria where those in activity one do not all select the same care level. If there are different care levels being chosen simultaneously in equilibrium, it must be true that the individuals in activity one are indifferent between the different care levels. This situation arises, for a care level y in activity two, if the due care standard is equal to $\hat{d}_x(y)$. In this case those in activity one are indifferent between d_x and $x_2(y)$. Of course, this situation can only arise when those in activity two are not negligent, for only then is there an incentive to reach the due care level in activity one.

Let us assume that some fraction, α, of those in activity one are at d_x, while $(1 - \alpha)$ of those in activity one are at $x_2(y)$. If we assume that expected accident costs are linear in the numbers engaged in the other activity and additive across participants selecting different care levels, we can write expected utility in activity two as

$$V = \begin{cases} \alpha B(d_x,y) + (1-\alpha)B(x_2,y) - \alpha C(d_x,y) & : y \geqq d_y \\ \alpha B(d_x,y) + (1-\alpha)B(x_2,y) - \alpha C(d_x,y) & \\ \qquad - (1-\alpha)C(x_2,y) \quad : y < d_y. & \end{cases} \tag{43}$$

(We assume that those at x_2 are negligent.) Let us denote by $y_3(\alpha,x_2,d_x)$ and $y_4(\alpha,x_2,d_x)$ the optimizing levels of care for the two utility functions. Thus y_3 satisfies

$$\alpha B_y(d_x,y_3) + (1-\alpha)B_y(x_2,y_3) - \alpha C_y(d_x,y_3) = 0. \tag{44}$$

In equilibrium of this type, those in activity two must be choosing either d_y or y_3 (when it exceeds d_y).

Before considering equilibrium, we shall examine some properties of y_3. Let us note the ambiguity in the response of y_3 to α:

$$\frac{\partial y_3}{\partial \alpha} = - \frac{B_y(d_x,y_3) - B_y(x_2,y_3) - C_y(d_x,y_3)}{\alpha B_{yy}(d_x,y_3) - (1-\alpha)B_{yy}(x_2,y_3) - \alpha C_{yy}(d_x,y_3)}. \tag{45}$$

Since B_{xy} is negative and d_x exceeds x_2, the first two terms in the numerator sum to a negative number. $-C_y$ however is positive. This ambiguity is natural since an increase in α decreases unrecognized accident costs (and their marginal response to care) but increases the recognized accident costs which are borne in activity two (and their marginal response to care). Let us note, however, that

$$\begin{aligned} y_3(1,x_2,d_x) &= y_2(d_x) \\ y_3(0,x_2,d_x) &= y_1(x_2). \end{aligned} \tag{46}$$

We assume that y_3 is monotonic in α so that y_3 for intermediate values of α lies in the range defined by the values of y_3 at the endpoints $\alpha = 0$ and $\alpha = 1$. The elements which determine the relative sizes of these two values are the same elements which determine whether there is a gap or overlap between regions II and IV.

Let us start by considering nonuniform equilibria where those in activity two are precisely at the due care standard. We know that they will make this choice provided d_y lies in the interval

$$y_3(\alpha,x_2,d_x) \leqq d_y \leqq \hat{d}_y(\alpha,x_2,d_x). \tag{47}$$

As above, $\hat{d}_y$ is defined by equality of the nonnegligent utility level and the optimal negligent utility level:

$$\begin{aligned} &\alpha B(d_x,\hat{d}_y) + (1-\alpha)B(x_2,\hat{d}_y) - \alpha C(d_x,\hat{d}_y) = \\ &\alpha B(d_x,y_4) + 1 - \alpha)B(x_2,y_4) - \alpha C(d_x,y_4) - (1-\alpha)C(x_2,y_4). \end{aligned} \tag{48}$$

Given a value of d_y satisfying the inequality (47) for some value of α, we have an equilibrium with the matching due care standard in activity one

$$d_x = \hat{d}_x(d_y). \tag{49}$$

The question then is the range of values of d_y which satisfy the inequality (47) for some α in the unit interval when evaluated at $x_2(d_y)$ and $\hat{d}_x(d_y)$. For each α we get a line segment of values of d_y satisfying the condition. Taking α to be zero, the inequality (47) is the same as that which partially defines region III. Thus we get this type of equilibrium for pairs of due care standards lying on the curve $d_y = \hat{d}_x(d_y)$ from $\bar{d}_y$ to y^{00}. Taking α to be one, y_3 and $\hat{d}_y$ coincide, and we have the point defined by the intersection of $y_2(x)$ and $\hat{d}_x(y)$, which is the northeast corner of region II. By continuity in α we have a locus of equilibria along the curve $d_x = \hat{d}_x(d_y)$ at least from $\bar{d}_y$ to the lower point of y^{00} and $y_2(d_x)$. This region is only a line, but serves as the introduction to the region where those in activity two are at y_3.

In the case where d_y is precisely equal to $y_3(\alpha,x_2,d_x)$ we have an equilibrium of the type described (since d_y is at least as large as y_3). With any decrease in d_y, those in activity two remain at y_3 and the equilibrium is unchanged. Thus we have a region, call it region V defined by the conditions

$$\begin{aligned} d_y &\leqq y_3(\alpha,x_2,d_x) \qquad \text{for some } \alpha \\ d_x &= \hat{d}_x(y_3), \quad \text{with} \quad x_2 = x_2(y_3). \end{aligned} \tag{50}$$

The question for defining this region is to determine the part of the curve $d_x = \hat{d}_x(d_y)$ for which d_y is precisely equal to y_3 for some value of α. That is, we want the values of d_y satisfying

$$d_y - y_3[\alpha,x_2(d_y),\hat{d}_x(d_y)] \qquad \text{for some } \alpha. \tag{51}$$

By the hypothesized monotonicity[18] of y_3 in α and continuity, d_y satisfies (51) if and only if it lies in the interval defined by $y_3[0,x_2(d_y),\hat{d}_x(d_y)]$ and $y_3[1,x_2(d_y),\hat{d}_x(d_y)]$. But this is the interval defined by $y_2[\hat{d}_x(d_y)]$ and $y_1[x_2(d_y)]$. Given the assumed relative slope conditions, this condition will be satisfied for the values of d_y lying between the values defined by

$$d_y = y_2[\hat{d}_x(d_y)] \tag{52}$$

$$d_y = y_1[x_2(d_y)]. \tag{53}$$

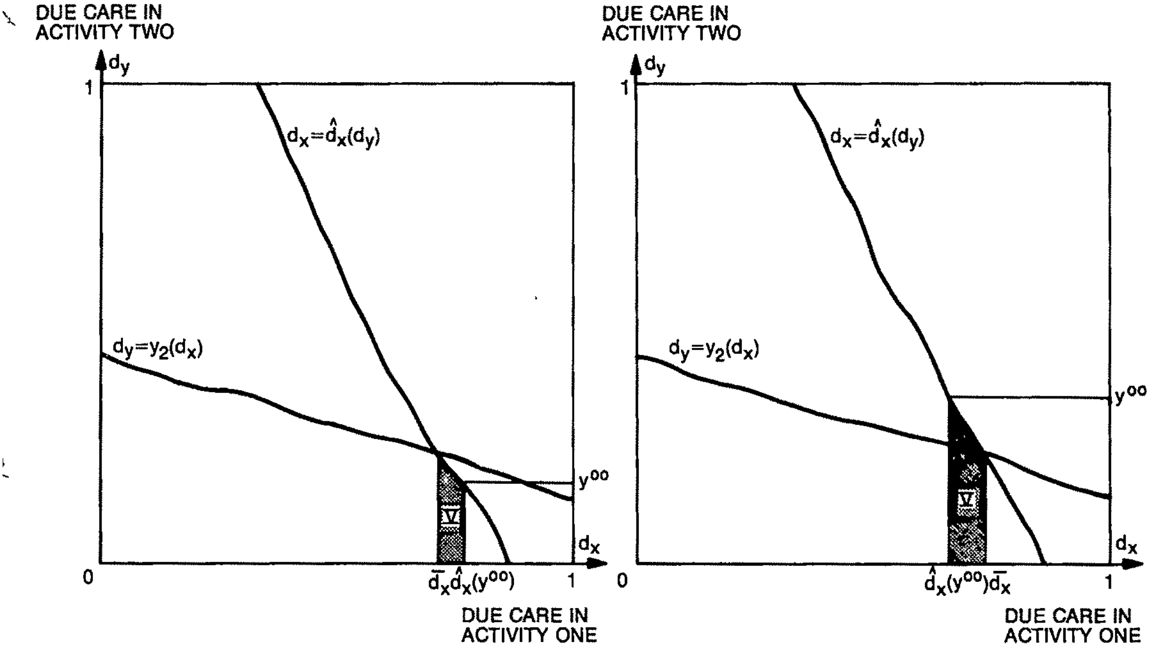
But these values are precisely $y_2(\bar{d}_x)$ and y^{00}. Thus, as is shown in Figure 25, region V fills the area between the borders of regions II and IV whether there is a gap or an overlap. Figure 26 shows a value of d_y which satisfies these conditions. Summarizing, we get

---

[18] Without monotonicity, region V might be larger than described here.

FIGURE 25

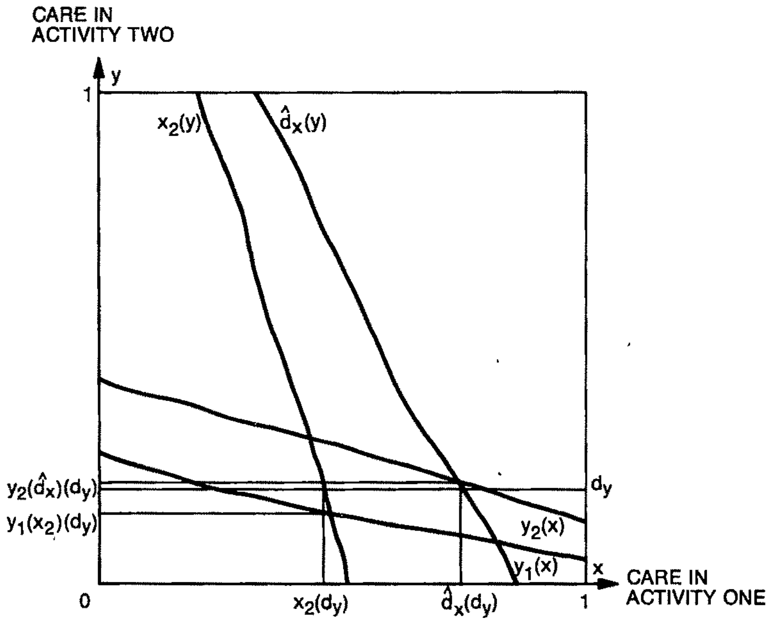
REGION OF TYPE V EQUILIBRIUM



nonuniform equilibria in region V and along the curve $d_x = \hat{d}_x(d_y)$. In the latter case α can be changed without altering equilibrium care values. This is not true in region V.

FIGURE 26

AN EQUILIBRIUM OF TYPE V



☐ **Nonuniformity in activity two.** In a parallel fashion we can consider the case where a fraction, β, of those in activity two select a care level equal to the due care standard, d_y, while the remainder, $1-\beta$, select the best negligent point, y_2. These positions must have the same utility. In activity one, then, utility satisfies

$$U=\begin{cases}\beta A(x,d_y)+(1-\beta)A(x,y_2) & :x\geqq d_x\\ \beta A(x,d_y)+(1-\beta)A(x,y_2)-\beta C(x,d_y) & :x<d_x.\end{cases}\tag{54}$$

Let us denote the optimizing care levels for the two functions by $x_3(\beta,y_2,d_y)$ and $x_4(\beta,y_2,d_y)$. Also let us denote by $\hat{d}_x(\beta,y_2,d_y)$ the care level which gives the same utility as the best negligent point

$$\beta A(\hat{d}_x,d_y)+(1-\beta)A(\hat{d}_x,y_2)= \beta A(x_4,d_y)+(1-\beta)A(x_4,y_2)-\beta C(x_4,d_y).\tag{55}$$

For those in activity two to have the same expected utility at y_2 as at d_y (with these values different), those in activity one must be negligent. Thus the conditions defining this type of equilibrium, region VI, are:

$$\text{for some } \beta \qquad d_x\geqq \hat{d}_x[\beta,y_2(x_4),d_y]\tag{56}$$

$$d_y=\hat{d}_y(x_4)\tag{57}$$

$$\text{where} \qquad x_4=x_4[\beta,y_2(x_4),d_y].\tag{58}$$

We have two separate questions to ask. First, for what values of d_y can we find a value of β giving a simultaneous solution to (57) and (58). Secondly, for the set of β giving equilibrium for each such value of d_y, what is the minimal value of d_x satisfying (56).

In a parallel fashion to that of the previous section we can show that we can find such an equilibrium if and only if d_y lies in the range defined by $\hat{d}_y(x^0)$ and $\bar{d}_y$. Thus the values of d_y for this type of equilibrium must lie between the upper boundary of region III and the lower boundary of region I (see Figure 17) independent of which of these values is the greater.

It remains to enquire about the values of d_x which will satisfy (56) for these values of d_y and β. Assuming that x_4 is monotonic in β there is a unique β, call it $\beta(d_y)$, such that (57) and (58) are satisfied. The equation for the boundary of region VI can be written

$$d_x=\hat{d}_x[\beta(d_y),y_2(x_4),d_y]\equiv f(d_y) \quad \text{for} \quad x_4=\hat{d}_y^{-1}(d_y).\tag{59}$$

One can check that $f(d_y)$ passes through the corners of regions I and III. When there is a gap between regions I and III [i.e., $\hat{d}_y(x^0)>\bar{d}_y$], $f(d_y)$ lies to the right of $\hat{d}_x(d_y)$. Thus, in this case the gap is completely filled. All three cases are shown in Figure 27.

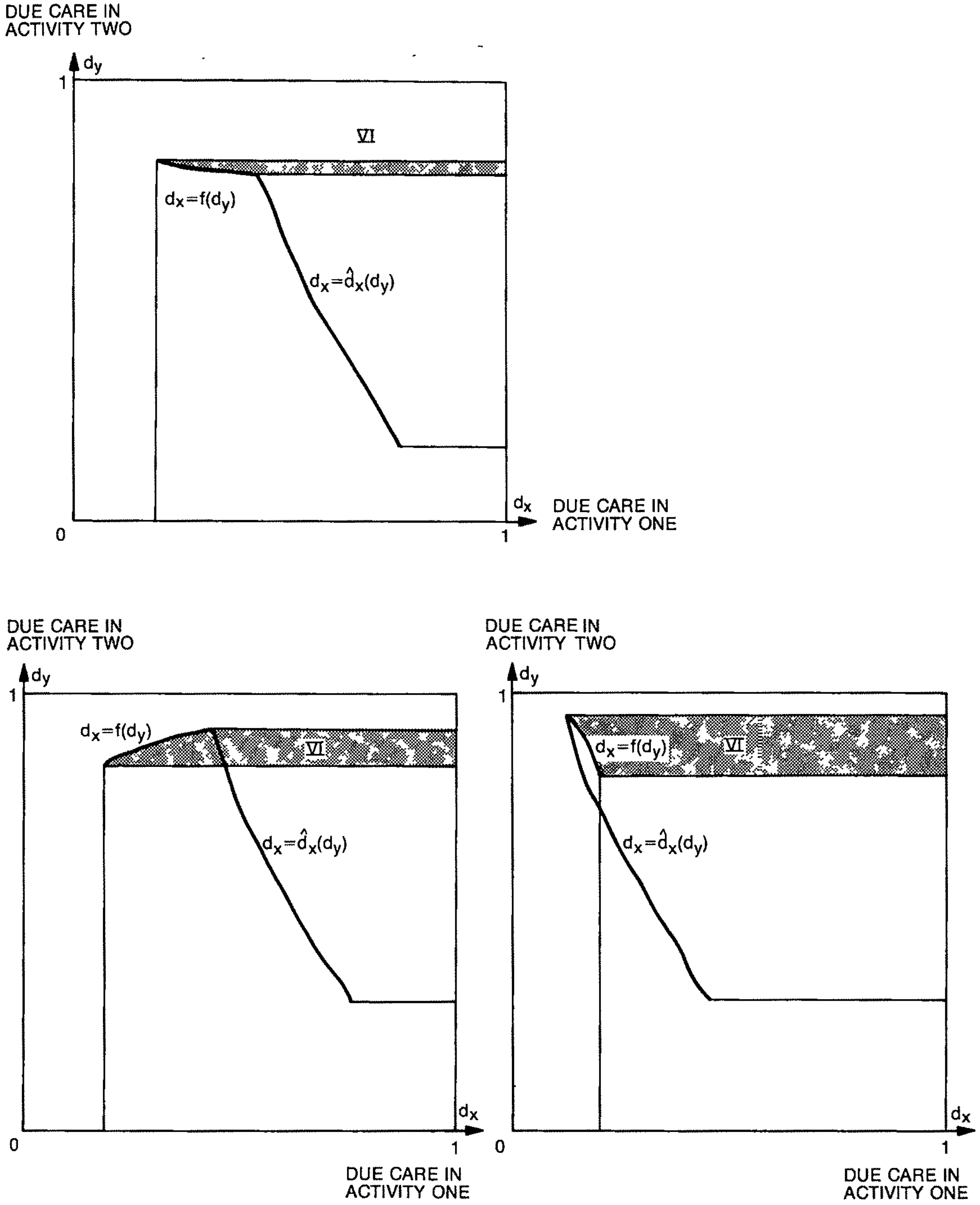
☐ **Existence of equilibrium.** Examining the map of regions for the case without overlaps of regions, we have the situation depicted by Figure 28. There thus remains a central section whose boundaries are

$$d_x=x^0, d_x=f(d_y), d_x=\hat{d}_x(d_y), d_y=y_2(d_x).\tag{60}$$

(With overlaps we would add the curves $\hat{d}_x=d_x(y^{00})$, $d_y=\hat{d}_y(x^0)$ to the list of boundaries.) It is clear that a central region

FIGURE 27

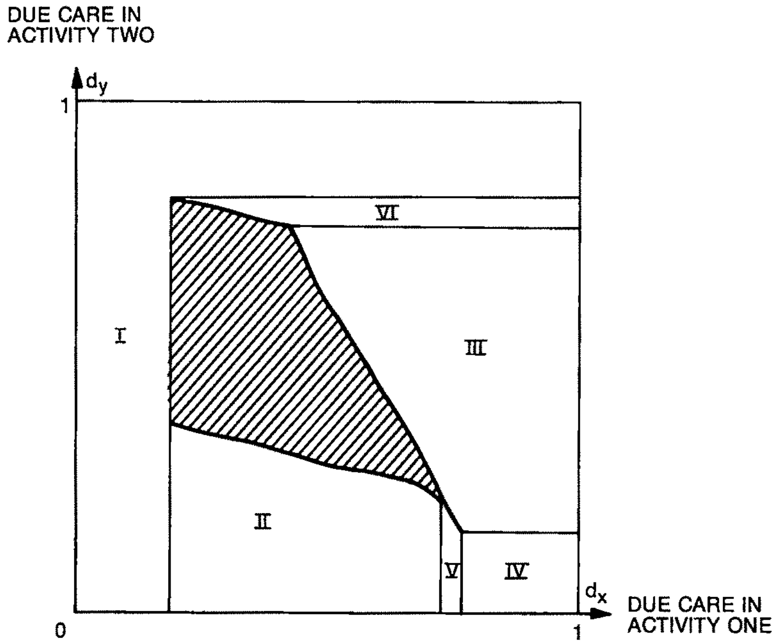
REGION OF TYPE VI EQUILIBRIUM



necessarily exists under the assumptions we have made about the slopes of the different curves. Rather than exploring the full size of the region with nonuniformity in both activities, we have proved the existence of an equilibrium for any pair of due care standards in the central region. Thus there always exists an equilibrium. Since the proof is not particularly instructive, it is omitted here, and we shall simply state the definition of this type of equilibrium. Let us begin with the definitions we shall need:

FIGURE 28

EXAMPLE OF GAP IN EQUILIBRIA OF TYPES I-VI



$$\begin{aligned}
&x_4(\beta,y,d_y) \text{ maximizes } \beta A(x,d_y) + (1-\beta)A(x,y) - \beta C(x,d_y)\\
&y_4(\alpha,x,d_x) \text{ maximizes } \alpha B(d_x,y) + (1-\alpha)B(x,y) - \alpha C(d_x,y)\\
&\qquad\qquad\qquad - (1-\alpha)C(x,y)\\
&\hat{d}_x(\beta,y,d_y) \text{ is the solution which is at least as large as}\\
&\quad x_4(\beta,y,d_y) \text{ to } \beta A(\hat{d}_x,d_y) + (1-\beta)A(\hat{d}_x,y)\\
&\qquad = \beta A(x_4,d_y) + (1-\beta)A(x_4,y) - \beta C(x_4,d_y)\\
&\hat{d}_y(\alpha,x,d_x) \text{ is the solution which is at least as large as}\\
&\quad y_4(\alpha,x,d_x) \text{ to } \alpha B(d_x,\hat{d}_y) + (1-\alpha)B(x,\hat{d}_y) - \alpha C(d_x,\hat{d}_y) =\\
&\quad \alpha B(d_x,y_4) + (1-\alpha)B(x,y_4) - \alpha C(d_x,y_4) - (1-\alpha)C(x,y_4).
\end{aligned} \tag{61}$$

We shall have an equilibrium of type VII when we can find six values—α, β, x, y, d_x, d_y—with $x < d_x$ and $y < d_y$, such that

$$\begin{aligned}
x &= x_4(\beta,y,d_y)\\
y &= y_4(\alpha,x,d_x)\\
d_x &= \hat{d}_x(\beta,y,d_y)\\
d_y &= \hat{d}_y(\alpha,x,d_x).
\end{aligned} \tag{62}$$

By construction, individuals in activity one are indifferent between the care levels x and d_x. Similarly, those in activity two are indifferent between y and d_y. Thus the populations in each activity can be divided arbitrarily between the different care levels. Thus we shall have an equilibrium for any pair of values (α,β) (both in the unit interval) for which four values can be found simultaneously satisfying (62). In this region a change in either due care standard leads to a change in equilibrium levels of care.[19]

---

[19] The set of equilibria of this type has an interesting difference from the equilibria of the other types. Consider the reversed system where the person in activity one pays the recognized costs of the person in activity two unless the former is not negligent *and* the latter is negligent. Obviously, this change often results in a changed equilibrium for given due care standards. However, apart from region VII, this reversed system is congruent with the

8. Potential liability in activity two

■ The analysis above was based on the assumption that those in activity one might bear some of the costs of those in activity two, but those in activity two would not bear any of the costs of those in activity one. In many circumstances the situation is closer to symmetric in that costs may be transferred in either direction by the courts, depending on the pattern of negligence. Let us briefly consider possible uniform equilibrium configurations without analyzing the determinants of the type of equilibrium which will occur. A key element of the structure is the same definition of the care standard for negligence and contributory negligence by one party.[20] Above we had four possible uniform equilibria—internal solutions with one party or the other bearing recognized costs, and corner solutions with one party or the other just at the due care level. Now there are two types of recognized costs, C^1 and C^2 borne by those in activities one and two, respectively, in the absence of a liability system. Thus, there are three possible internal solutions, as those in activity one bear no recognized costs, C^1, or $C^1 + C^2$. Also we now have the possibility that both parties will be at the due care standard. Let us sketch this structure briefly.

In the absence of liability the net utility functions are

$$A(x,y) - C^1(x,y) \tag{63}$$

$$B(x,y) - C^2(x,y). \tag{64}$$

Since there are two transferable costs, we recognize three internal solutions

$$\begin{aligned} &x_1 \text{ maximizes } A \\ &x_2 \text{ maximizes } A - C^1 \\ &x_3 \text{ maximizes } A - C^1 - C^2. \end{aligned} \tag{65}$$

Similarly, we have

$$\begin{aligned} &y_1 \text{ maximizes } B \\ &y_2 \text{ maximizes } B - C^2 \\ &y_3 \text{ maximizes } B - C^1 - C^2. \end{aligned} \tag{66}$$

Possible equilibrium configurations are (x_1,y_3), (x_2,y_2), (x_3,y_1). In the first case those in activity one are naturally nonnegligent ($d_x < x_1$) while those in activity two do not find it worthwhile to be nonnegligent. In the second case either both parties are naturally not negligent or neither party finds it worthwhile to be nonnegligent. The third case is the obverse of the first.

If those in activity one are precisely at the due care standard, those in activity two might be negligent (and at y_3), nonnegligent because of just reaching the due care standard, or naturally nonnegligent and at y_2. We will not have an equilibrium with those in

---

traditional system in the following sense. For any equilibrium in regions I-VI achievable under the traditional system, there exists a pair of due care standards resulting in the same equilibrium care levels under the reversed system. In region VII the change in systems alters the definitions of x_4, y_4, $\hat{d}_x$, and $\hat{d}_y$ and so may give rise to a different set of equilibria.

[20] Prosser in [6], p. 418, argues that this is the situation in the United States. If this were not true, each party might successfully sue the other. To analyze this situation we would need to consider four due care standards rather than two.

activity two at y_1, for those in activity one, being nonnegligent could not be bearing C^2. Naturally all of these cases have their mirror images. In addition there are nonuniform equilibria which could be considered.

9. Stochastic model

■ The models above assumed both that each individual had control over his own accident-related behavior and that the courts could accurately measure the decisions of individuals. Dropping either of these elements adds an additional stochastic element to that of the occurrence of accidents. In many situations it is reasonable to model the choices of individuals as a selection among different probability distributions of actual behavior over time. For example, driving habits generate a probability distribution of drifting into the wrong lane or general safety policies of firms affect the probabilities of careless behavior by employees. Since I want to focus on the role of the contributory negligence standard, I shall consider a specific model of recognized costs depending on stochastic elements in activity two, while preserving the determinate model in activity one.[21]

Let us denote by a the level of care in activity two. (We assume both that a affects probabilities and that it is the instantaneous behavior, a, which is measured by a court which is making a contributory negligence evaluation.) We assume that a is uniformly distributed between 0 and y where y remains the control variable of the individual and is referred to as precaution; that there are no unrecognized costs in activity two, so that utility in the absence of accidents can be written as $B(y)$; that in the event of an accident, there are recognized costs, C, which do not vary with the decisions of the parties; and that the probability of an accident given care levels a and x (and the number of persons in activity one) is $(1 - a)(1 - x)$. Since the probability of taking care a is y^{-1} we have total expected recognizable costs

$$C(x,y) = C\int_0^y y^{-1}(1-x)(1-a)da = C(1-x)(1-\tfrac{1}{2}y). \tag{67}$$

Since we shall focus on the role of contributory negligence, we shall consider a due care standard in activity one that is prohibitively high, $d_x = 1$ (or, equivalently, that there is a strict liability with a contributory negligence standard.)

We shall denote the due care standard in activity two by d_a. Then, provided that d_a is not so high that everyone in activity two chooses to be contributorily negligent all of the time, we have the two utility functions

$$U = A(x,y) - C\int_{d_a}^y y^{-1}(1-x)(1-a)da \tag{68}$$

$$V = B(y) - C\int_0^{d_a} y^{-1}(1-x)(1-a)da. \tag{69}$$

---

[21] An alternative interpretation of this model is that individual decisions are determinate, but generate a distribution of measurements by the court's attempting to determine behavior at the time of the accident.

Where we have continued to write expected utility in activity one net of costs as $A(x,y)$. Performing the integrations we have

$$U = A(x,y) - C(1-x)y^{-1}[y - d_a - \tfrac{1}{2}(y^2 - d_a^2)] \quad (70)$$

$$V = B(y) - C(1-x)y^{-1}(d_a - \tfrac{1}{2}d_a^2). \quad (71)$$

Considering the optimal level of care in activity one, we have $x_5(y,d_a)$ as the solution to

$$A_x(x_5,y) + Cy^{-1}[y - d_a - \tfrac{1}{2}(y^2 - d_a^2)] = 0. \quad (72)$$

Calculating the response of the supply of care to its two determinants by implicit differentiation of (79), we have

$$\frac{\partial x_5}{\partial y} = -\frac{A_{xy} - \frac{1}{2}C + Cy^{-2}(d_a - \frac{1}{2}d_a^2)}{A_{rx}} \quad (73)$$

$$\frac{\partial x_5}{\partial d_a} = \frac{Cy^{-1}(1 - \frac{1}{2}d_a)}{A_{xx}} < 0. \quad (74)$$

Despite the assumption that $A_{xy} < 0$, the sign of $\partial x_5/\partial y$ is not clear depending on the size of y relative to d_a. For sufficiently small y it must be positive, while for sufficiently large y it must be negative. Without further justification, we assume a single sign change. To see the source of this ambiguity we notice that an increase in y decreases accident probability but increases the fraction of accidents in which the participant is not contributorily negligent. These conflicting forces affect the expected gain from more care in activity one.

In activity two, we have the supply of precaution $y_5(x,d_a)$ determined by

$$B_y(y_5) + C(1-x)y_5^{-2}(d_a - \tfrac{1}{2}d_a^2) = 0. \quad (75)$$

Thus, by implicit differentiation we have

$$\frac{\partial y_5}{\partial x} = \frac{Cy^{-2}(d_a - \frac{1}{2}d_a^2)}{B_{yy}(y) - 2C(1-x)y^{-3}(d_a - \frac{1}{2}d_a^2)} < 0 \quad (76)$$

$$\frac{\partial y_5}{\partial d_a} = \frac{-C(1-x)y^{-2}(1-d_a)}{B_{yy}(y) - 2C(1-x)y^{-3}(d_a - \frac{1}{2}d_a^2)} > 0. \quad (77)$$

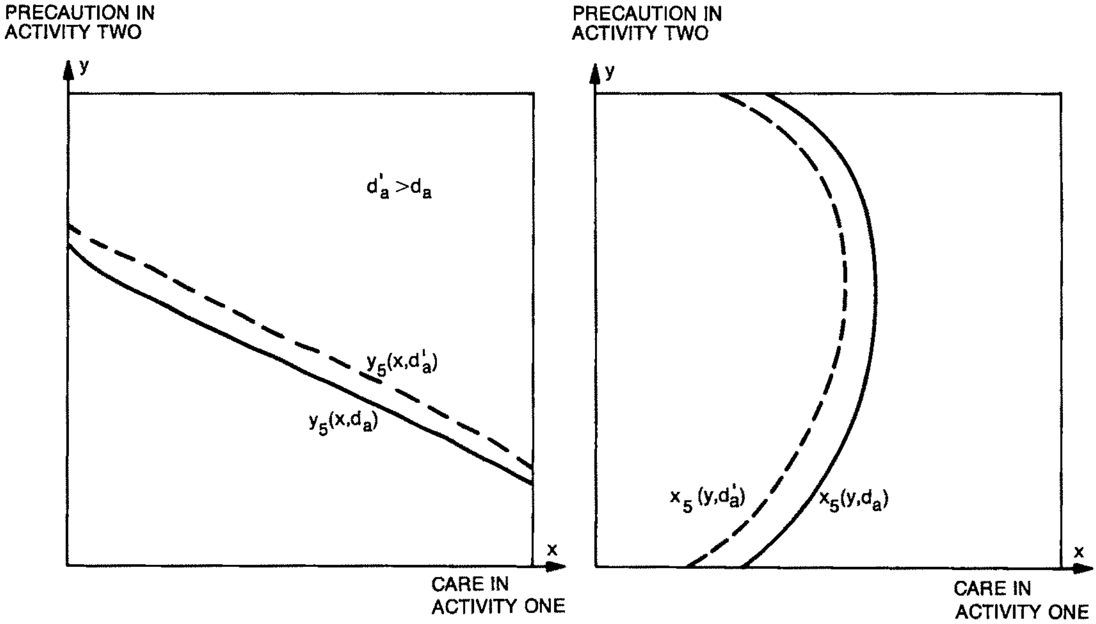
Thus, an increase in the contributory negligence standard increases the supply of precaution and decreases the supply of care in activity one.

To determine the general equilibrium impact of an increase in the contributory negligence standard on the two decisions, we need to combine the responses of the two individuals to the standard with their responses to each other. As above, we assume a unique equilibrium, with choice occurring in the interior of the choice set. The individual responses to an increase in d_a are shown in Figure 29, while the two equilibrium possibilities under the uniqueness assumption are shown in Figure 30.

As the diagrams show, the equilibrium quantity of precaution definitely increases with the contributory negligence standard. The equilibrium quantity of care will fall when care decreases with precaution. When care increases with precaution, however, a suf-

FIGURE 29

INDIVIDUAL RESPONSES TO AN INCREASE IN THE DUE CARE STANDARD IN ACTIVITY TWO



ficiently large response of precaution to the standard may, in its indirect effect, overcome the direct tendency of care to fall with the standard, resulting in an increase in both. The algebra of these diagrams can be obtained by simultaneously differentiating the two first-order conditions:

$$\frac{dx_5}{dd_a} = D^{-1}\left(\frac{\partial x_5}{\partial d_a} + \frac{\partial x_5}{\partial y}\frac{\partial y_5}{\partial d_a}\right) \tag{78}$$

$$\frac{dy_5}{dd_a} = D^{-1}\left(\frac{\partial y_5}{\partial d_a} + \frac{\partial y_5}{\partial x}\frac{\partial x_5}{\partial d_a}\right), \tag{79}$$

where

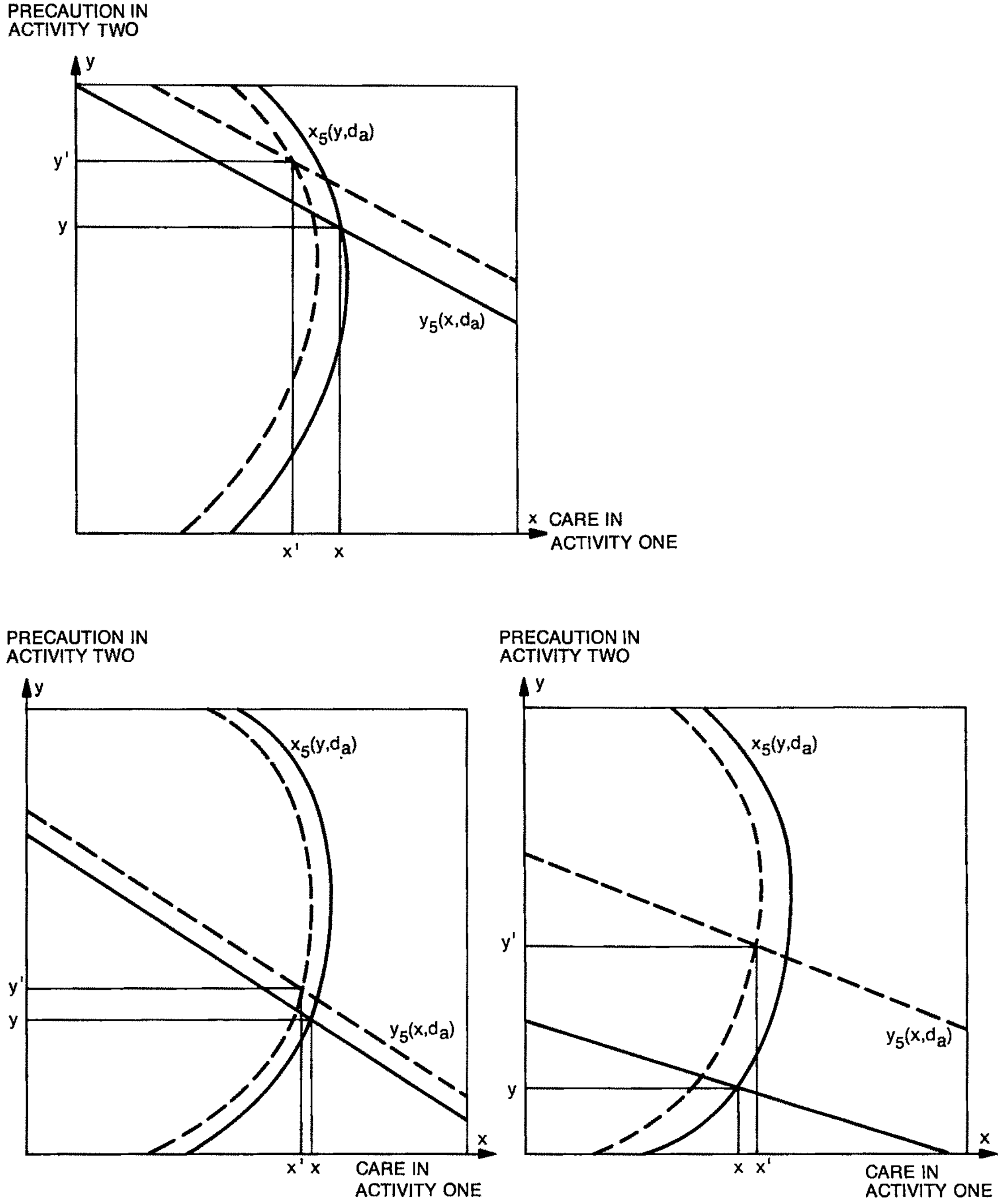
$$D = 1 + \frac{\partial x_5}{\partial y}\frac{\partial y_5}{\partial x}. \tag{80}$$

D is positive by the stability condition that necessarily follows from the uniqueness of equilibrium. Thus $\frac{dy}{dd_a}$ is necessarily positive in this range where y exceeds d_a.

In considering the effect of different contributory negligence standards on the sum of utilities, there are two general propositions. First, it is not possible to achieve the efficient solution as an equilibrium in this model with strict liability in activity one. Second, the sum of utilities is increasing in the due care standard for very low due care standards (i.e., sufficiently close to zero). In addition, one could examine the equation for the due care standard which maximizes the sum of utilities over the available values. To see the first point assume that it is possible to induce the optimal choice of y for some due care standard. Since this standard is not zero, those in activity two are sometimes contributorily negligent. Thus, those

FIGURE 30

CHANGE IN EQUILIBRIUM FROM AN INCREASE IN THE DUE CARE STANDARD IN ACTIVITY TWO



in activity one are not bearing all the marginal costs from their last decrease in care. As a result, they do not have an incentive to pick the correct care level. I have not examined the circumstances in which it is possible to achieve the efficient solution by setting the due care standard in activity one precisely at the desired care level, while simultaneously choosing an appropriate due care standard in activity two.

To examine the second proposition, let us examine the sum of utilities as a function of the due care standard:

$$S(d_a) = A(x,y) + B(y) - C(1-x)(1-\tfrac{1}{2}y). \tag{81}$$

Differentiating, we have

$$S'(d_a) = (A_x + C(1-\tfrac{1}{2}y))\frac{dx}{dd_a} + (A_y + B_y + \tfrac{1}{2}C(1-x))\frac{dy}{dd_a}. \tag{82}$$

Making use of the first-order conditions, (72) and (75), we can write this as

$$S'(d_a) = Cy^{-1}(d_a - \tfrac{1}{2}d_a{}^2)\frac{dx}{dd_a} + [A_y + \tfrac{1}{2}C(1-x) - Cy^{-2}(1-x)(d_a - \tfrac{1}{2}d_a{}^2)]\frac{dy}{dd_a}. \tag{83}$$

When d_a is equal to zero, this expression is necessarily positive.

To examine the maximizing due care standard, let us assume that A_y is zero, so that all costs are recognized by the legal system. Setting $S'(d^*{}_a)$ equal to zero, we have the equation

$$(d_a^* - \tfrac{1}{2}d_a^{*2}) = \frac{-\tfrac{1}{2}C(1-x)\dfrac{dy_5}{dd_a}}{Cy^{-1}\dfrac{dx_5}{dd_a} - Cy^{-2}(1-x)\dfrac{dy_5}{dd_a}} \tag{84}$$

$$= \frac{\tfrac{1}{2}y^2}{\dfrac{\partial y_5}{\partial d_a}\left[1 + y(1-x)\dfrac{\partial x_5}{\partial y}\right] - \dfrac{\partial x_5}{\partial d_a}\left[\dfrac{\partial y_5}{\partial x} + y(1-x)\right]}.$$

Whether anything useful can be done with this equation is not clear.

10. Concluding remarks

■ Obviously any model of accidents and tort law which has no risk aversion, no insurance, no errors of perception, and no costs or errors of measurement is not directly applicable for policy purposes. But it might serve as a building block in the eventual development of models that consider simultaneously the problems of decisions in equilibrium, the distribution of the bearing of accident costs, and the administrative costs of running the system, along with the other social tools for altering incentives. It may also help in the development of the understanding of complicated issues, such as whether to have a contributory negligence doctrine where the doctrine tends to encourage more care on the part of some at a cost of discouraging care on the part of others.

References

1. Brown, J. P. "Towards an Economic Theory of Liability." *Journal of Legal Studies*, Vol. 2 (1973).
2. Calabresi, G. *Costs of Accidents.* New Haven: Yale Univ. Press, 1970.
3. Diamond, P. "Single Activity Accidents." *Journal of Legal Studies*, Vol. 3, No. 1 (January 1974), pp. 107–162.

4. LEFF, A. "Injury, Ignorance, and Spite." *Yale Law Journal*, Vol. 80, No. 1 (October 1970), pp. 1–46.
5. POSNER, R. "A Theory of Negligence." *Journal of Legal Studies*, Vol. 1 (1972).
6. PROSSER, W. *Law of Torts*, 4th ed. St. Paul, Minn.: West Publishing Co., 1971.

The choice between capitalizing and expensing under rate regulation

Peter B. Linhart

Treasury Department
American Telephone and Telegraph Company

Joel L. Lebowitz

Belfer Graduate School of Science
Yeshiva University

and

Frank W. Sinden

Bell Telephone Laboratories, Murray Hill

In this paper we give a mathematical analysis of some of the consequences, over time, of the decision to capitalize or to expense. Both regulated and unregulated firms are considered. It is shown that for a regulated firm, in contrast to an unregulated one, this decision does have an impact on the customers, who should rationally prefer either capitalization or expensing, depending on their discount rate. It is shown that the attitudes of rational owners will also depend on their discount rate. The point of view of the tax collector is also considered.

1. Introduction

■ This paper examines the relative effects on revenue, taxes, earnings, outside capital requirements, etc., of "capitalizing" certain items of expenditure, that is, raising capital to pay for them, as opposed to "expensing" them, paying for them outright.

It is assumed that these expenditures will be made in any case, so that the choice between capitalization and expensing is an *accounting* choice,[1] which does, however, have economic conse-

---

Peter B. Linhart received the B.A. from Princeton University (1948), the M.A. from the University of California (1950), and the Ph.D. from Columbia University (1963). He is currently working on corporate financial problems.

Joel L. Lebowitz received the B.S., the M.S., and the Ph.D. from Syracuse University (1952, 1955, and 1956, respectively). He is currently investigating exact results in statistical mechanics.

Frank W. Sinden received the B.S. from the University of Chicago (1948) and the Dr.Math., Sci. from the Swiss Federal Institute of Technology, Zurich (1954). Currently he is interested in economics and environmental problems.

The authors are indebted to M. B. Wilk of AT&T for helpful suggestions and comments. We are also grateful to J. M. Bank, B. Elias, H. A. Eakin, and J. A. Fischer, all of AT&T, and to B. L. Behler of Bell Telephone Laboratories, for valuable help.

[1] Thus we are *not* discussing the economic choice between labor and capital as factors of production, discussed by Averch and Johnson [1], and in many subsequent papers. See, in particular, Bailey and Malone [2].

quences, as we shall see. It should also be emphasized that the items under consideration are those for which such an option is allowed by the tax collectors, the accounting profession, and the regulatory body, if any—for example, items of furniture costing less than $100.

The difference between capitalizing and expensing is as follows. If an item is capitalized, it is paid for out of future revenues, but at the cost of providing a return to the suppliers of capital; capitalization spreads out expenditures into the future, and smooths them. If an item is expensed, it is paid for out of current revenues. One of the essential determinants of such a choice therefore involves a comparison between the cost of capital and the time value of money, as experienced by the various interested parties. Another determinant of the choice will be its effect on outside capital requirements. The interested parties include the customers, the owners, the managers, the regulators, and the tax collectors; the interests of these groups are in part conflicting.

To determine an "optimal" amount of capitalization, it would be necessary to balance off its effects on one group against those on another. We shall not attempt to do that in this paper. Instead, we shall calculate the effect over time of capitalizing in place of expensing on each of a number of variables of interest to the various groups, and leave to the reader any balancing of interests among them.

At this point we should make clear the distinction between an unregulated and a regulated firm. By an *unregulated* firm we mean one whose revenues are determined entirely by the market place, and are therefore independent of the capitalization-expensing decision. The customers of such a firm are clearly indifferent to this decision. For a *regulated* firm (to which this paper is primarily devoted), a public commission sets prices for the firm's products in such a way that an allowed rate of return—the ratio of net earnings to total capital—will be realized.[2] The customers of a regulated firm clearly are affected by the capitalization-expensing decision.

In reality, of course, regulation is intermittent[3] and not absolute. Thus a real regulated firm might be expected to lie between the two extreme cases described above. However, since our main point in this note is to present in simple mathematical form the direct, first-order effects which are caused by a switch from capitalization to expensing, we shall assume here that regulation is perfect. For the same reason, we shall neglect here the price-elasticity of demand, and assume that revenues and expenses (hence earnings) are predictable with complete certainty. In such a world of perfect certainty, it would follow that the cost of capital to the firm, expressed in its discount rate, as well as the rate of return on capital which the regulatory commission should permit, should be equal to the pure interest rate, i. Such certainty is, of course, not present in the real world and the difference between the pure interest rate, on the one hand, and, for example, the cost of equity, on the other, is one reflection of this uncertainty. The appropriate formulation of the exact form of the relationship between risk and expected return on investment is a question which has been much discussed.[4]

---

[2] For a discussion of price-setting in regulated firms, see Davis [4] and references therein.

[3] See Klevorick [6].

In simple cases the results may be seen easily by intuition. Consider, for example, a strictly regulated firm with no taxes. In this case the choice between expensing or capitalizing an item is, from the customers' point of view, equivalent to the choice between buying something outright and buying it on time. The "interest" the customers pay if the item is capitalized is equal to the company's rate of return on capital, and the payback schedule is the company's depreciation schedule for that item.

Complications such as taxes, however, can have a considerable effect on the results. In the presence of the corporate income tax, the customers' "interest" mentioned above becomes the *pre-tax* rate of return, which may be considerably higher than the commonly quoted after-tax rate of return set by the regulator.

If the distinction between book and tax depreciation (allowed by present law) is recognized, the mathematics becomes more complicated and additional terms appear in the results.

In order to simplify the formulas and circumvent tedious calculations, we make use of the *convolution* notation. This is particularly helpful in connection with present worths, which we use to compare different expenditure streams and hence to express the preferences of the various parties. If the meaning of the convolution is understood, the compact notation should also aid in the interpretation of results.

2. The model

■ The model on which our calculations are based is illustrated by the diagram of Figure 1, in which the symbols are also defined.[5] This diagram shows the cash flows per unit of (continuous) time, into, out of, and within the firm as here modeled. Revenues and outside capital flow in, and expenses, interest, etc. flow out.

Looking at the top node, we see that the total revenues, R, which the firm collects from its customers go after covering expenses, E, and book depreciation, D_B, into the pre-tax earnings, A, or

$$R = E + A + D_B. \tag{1}$$

The bottom node shows that the construction budget, C, is provided by the sum of the book depreciation, D_B, and the increase, per unit time, in the total capital of the firm, K. This capital consists of two parts: first that raised from the outside through the sale of stocks and bonds, denoted by K_O, and second that accumulated internally through the retention of earnings, denoted by K_I. Thus

$$C = D_B + \dot{K}_O + \dot{K}_I, \tag{2}$$

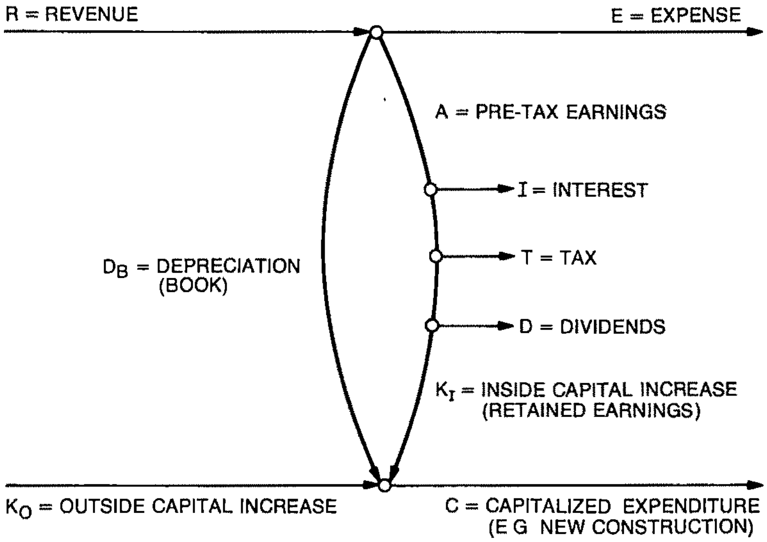
---

[4] See, for example, Cootner and Holland [3] and Sharpe [8].

[5] Some of the methods used here are similar to those of Linhart [7].

FIGURE 1

CASH FLOWS OF THE MODEL FIRM


R = REVENUE
E = EXPENSE
A = PRE-TAX EARNINGS
I = INTEREST
T = TAX
D = DIVIDENDS
D_B = DEPRECIATION (BOOK)
K_I = INSIDE CAPITAL INCREASE (RETAINED EARNINGS)
K_O = OUTSIDE CAPITAL INCREASE
C = CAPITALIZED EXPENDITURE (E G NEW CONSTRUCTION)

EACH ARROW REPRESENTS A (TIME-VARIABLE) FLOW WITH THE UNITS DOLLAR PER UNIT TIME. VARIABLES NOT SHOWN ARE DEFINED BELOW·

$K = K_O + K_I$ = TOTAL CAPITAL

δ = DEBT RATIO, i e., δK = debt, $(1-\delta)K$ = EQUITY

$a = A/K$ = PRE-TAX RATE OF RETURN

$\rho = (A-T)/K$ = POST-TAX RATE OF RETURN

τ = TAX RATE

$\imath = I/\delta K$ = INTEREST RATE

$d = D/(1-\delta)K$ = DIVIDEND RATE

X = UNIT NET PLANT (SEE TEXT)

SUBSCRIPTS B AND T REFER TO BOOK AND TAX DEPRECIATION

where the dot denotes the time derivative, e.g.,

$$\dot{K}_I = \frac{dK_I}{dt}.$$

$\dot{K}_I$ represents retained earnings, which are obtained from pre-tax earnings A, as shown in Figure 1, by the subtraction of taxes T, interest I, and dividends D,

$$\dot{K}_I = A - I - T - D. \qquad (3)$$

In addition, we have the tax equation

$$T = \tau(R - E - D_T - I), \qquad (4)$$

where D_T is depreciation as calculated for tax purposes. (Only income taxes are considered here.)

In the case of a regulated firm, there is a further accounting relation,

$$R = \rho K + T + D_B + E,$$

which expresses the fact that post-tax earnings, $A - T$, are a fixed fraction of capital K. This fraction, ρ, is the allowed rate of return set by the regulator. As mentioned earlier, we shall assume here, when dealing with a regulated firm, that this allowed rate of return is in fact realized.

Depreciation can be expressed in terms of a unit depreciation schedule and the sequence of capital outlays for new construction

$$D_B(t) = \int_0^t d_B(t-t')C(t')dt', \tag{5}$$

where $d_B(s)$ is the rate of depreciation (fraction/year) of a unit of plant of age s. (For tax depreciation the subscript T replaces B.) $C(t')$ is the rate of capitalized expenditure at time t'. The firm is established at time zero.

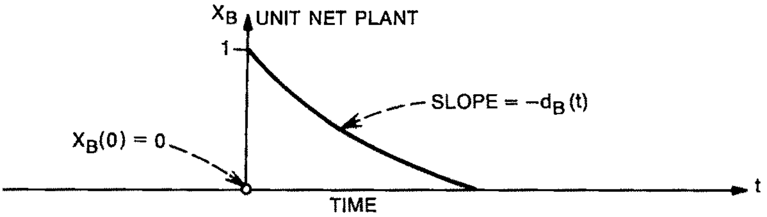
It will be convenient to use in place of the depreciation rate d_B the undepreciated fraction of a unit of plant remaining at age t or the *unit net plant*:

$$\begin{aligned} X_B(0) &= 0 \\ X_B(t) &= 1 - \int_0^t d_B(t')dt', \quad t > 0. \end{aligned} \tag{6}$$

The unit net plant jumps sharply from zero to one at the time $(t = 0)$ of acquisition and then falls gradually back to zero as the plant is depreciated. A typical time course for $X_B(t)$ is plotted in Figure 2.

FIGURE 2

SCHEMATIC FOR DEPRECIATION OF UNIT NET PLANT



Setting $X_B(0) = 0$ has the effect of including the acquisition jump within the nonnegative time domain, which is mathematically convenient.

3. Convolution notation

■ The integral in equation (5) is a *convolution* of the functions $d_B(t)$ and $C(t)$. Using a standard notation, equation (5) can be written

$$D_B = d_B * C. \tag{5a}$$

The convolution has many simple properties useful in calculation. For example, a change in the variable of integration shows that

$$d_B * C = C * d_B.$$

Other properties are given in the Appendix. The convolution may be thought of as superimposing weighted translates of a fixed function. The integral in equation (5), for example, superimposes the depreciation flows generated by various vintages of plant. That generated by vintage t' plant is the per unit depreciation schedule

$d_B(t)$, translated to start at t' and weighted by the amount $C(t')dt'$ of vintage t' plant that was installed.

A useful result states that total capital equals net plant, which is the sum of undepreciated plant from all past vintages:

$$K = X_B * C.$$

4. Application of the model

■ The total rate of expenditure required to run the business is $[E(t) + C(t)]$ (see Figure 1). $E(t)$ is the portion treated on the books as current expense; $C(t)$ is the portion treated as capital. Taking the physical operations and the sum $E(t) + C(t)$ as given, we wish to examine the effects of transferring, on the books, a stream of money $F(t)$ from the expense stream $E(t)$ to the capital outlay stream $C(t)$. $F(t)$ is an arbitrary function, zero until the time of transfer, and positive thereafter. $F(t)$ could be taken to be negative; we would then be considering a shift from capitalization to expensing. The reader can easily supply the corresponding sign changes in the results.

The subscript E (for "expensing") will indicate the value of a variable without the transfer, the subscript C (for "capitalization") its value with the transfer. Thus,

$$\begin{aligned} C_C &= C_E + F \\ E_C &= E_E - F. \end{aligned} \tag{7}$$

We consider first the case of a strictly regulated firm ($\rho_C = \rho_E$ given) with book and tax depreciation the same ($D_B = D_T$). In this case the pre-tax rate of return is unaffected (see Appendix):

$$a = a_C = a_E = (\rho - \tau i \delta)/(1 - \tau). \tag{8}$$

We shall now consider the effects of the decision on various groups.

☐ **Customers.** The customers are interested only in the effect on revenues, since these are what they pay. It is shown in the Appendix that

$$R_C - R_E = (aX_B - \dot{X}_B) * F. \tag{9}$$

The convolution on the right may be interpreted as a superposition of "loans" to the body of customers. The amount of the loan made at time t is equal to the amount $F(t)$ transferred to capital at that time. The "interest rate" paid by the customers is equal to the pre-tax of return a, and the unrepaid fraction is equal to the unit net plant X. Thus the term aX represents "interest" on the unrepaid balance, and $-\dot{X}$ represents repayments. Equation (6) shows that these follow the depreciation schedule. Since we included the initial jump to unity in our definition of X, the $-\dot{X}$ term includes the initial payment of the "loan" to the customers as well as the repayment back to the firm.

If F rises faster than e^{at}, then the current "loan" is always more than enough to cover current "interest," so that $R_C - R_E$ always remains negative. In this case capitalization yields a perpetual saving in revenue. This situation is analogous to a stock market bubble or a chain letter and is wonderful in that it generates something for nothing.

If we rule out this unlikely case, then $R_C - R_E$ must eventually become positive, so that early revenue savings are bought at the cost of a later revenue penalty. How are the customers to judge the value to them of such an ambiguous revenue differential? A simple answer is to suppose that they look at the present worth of $R_C - R_E$ with respect to some discount rate k related to the way they trade time against money.

If we let $PW_k\{Z\}$ stand for the present worth,[6] i.e.,

$$PW_k\{Z\} \equiv \int_0^\infty e^{-kt} Z(t)\,dt,$$

then it follows from equation (9) (see the Appendix) that

$$PW_k\{R_C - R_E\} = (a - k) PW_k\{X\} PW_k\{F\}. \qquad (10)$$

Under the present worth criterion, then, customers prefer capitalization if and only if the expression on the right is negative. Since both $PW_k\{X\}$ and $PW_k\{F\}$ are positive, this occurs if and only if the customers' discount rate k exceeds the pre-tax rate of return a.

Although this result is simple, it must be applied with some caution. It assumes, tacitly, that the body of customers is static, infinite-lived and possessed of a definable interest as a whole. In fact, the customers come and go in an endless stream and have only individual interests. Thus a revenue differential $(R_C - R_E)$ that varies in time necessarily trades off the interests of some customers against those of others. Specifically, capitalization favors early customers at the expense of later ones. The present worth, therefore, should be regarded at best as only a crude indicator of the customers' interest and should not be used without regard to the shape of $R_C - R_E$ over time.

More refined analysis would also recognize that the quantity of interest to the individual customer is not total revenue, but revenue per unit of service or perhaps revenue per customer. These quantities can, of course, be calculated, given the appropriate data.

Nevertheless, the time shape of $R_C - R_E$ and even its present worth give a simple and useful first cut into the very knotty problem of determining the customers' interest in capitalization versus expensing for a regulated firm.

☐ **Owners.** The customers do not constitute the only interested party. One concern of the owners (shareholders) and their representatives the managers—e.g., the treasurer—may well be the amount of new capital they must raise. It is shown in the Appendix that the effect on outside capital requirements of the transfer $F(t)$ from expense to capital is

$$\dot{K}_{OC} - \dot{K}_{OE} = -[(\rho - i\delta - d(1-\delta))X_B - \dot{X}_B]^{*}F. \quad (11)$$

This equation has the same form as the revenue equation (9). It therefore represents a superposition of "loans," but this time

---

[6] The present worth is a Laplace transform. For further discussion of properties of the Laplace transform, see Widder [9].

with a negative sign and with "interest" equal to $\rho - i\delta - d(1 - \delta)$. The latter quantity is readily recognized to be the *retained earnings relative to capital.* We will denote it by r:

$$r = \rho - i\delta - d(1 - \delta) = \frac{\dot{K}_I}{K}. \tag{12}$$

Then equation (11) can be rewritten:

$$\dot{K}_{OC} - \dot{K}_{OE} = -(rX_B - \dot{X}_B)*F. \tag{13}$$

If F grows less rapidly than e^{rt}, *then* $\dot{K}_{OC} - \dot{K}_{OE}$ eventually becomes negative [see argument following equation (9)]. But this means that capitalization eventually leads to a *savings* in outside capital requirements! This seemingly anomalous result is explained by noting that capitalization generates a stream of retained earnings which helps to satisfy capital requirements. If F grows slowly enough, the cumulative stream of retained earnings from past capitalizations may more than offset new capital requirements.

However, since r may be numerically small, the case in which F grows faster than e^{rt} is not unreasonable. In this case outside capital requirements are always greater under capitalization than under expensing, but (if $r > 0$) not so great as $F(t)$.

We now suppose that the quantity which the owner (or collectivity of owners) wishes to maximize is the net outflow to him. In terms of our model this is dividends minus the equity inflow. Denoting it by S, we have

$$S = D - (\dot{K}_O - \delta\dot{K}).$$

The effect on this outflow of the transfer of $F(t)$ from expense to capitalization is

$$S_C - S_E = (1 - \delta)\left(\frac{\rho - i\delta}{1 - \delta} X_B - \dot{X}_B\right) *F.$$

As before, this represents a superposition of "loans." The transfer to capitalization, in effect, causes the owner to lend money at the "interest rate"

$$\epsilon = \frac{\rho - i\delta}{1 - \delta}.$$

As always, the payback follows the depreciation schedule $\dot{X}$. The quantity ϵ is readily recognized to be the rate of return on equity. (Multiply it by equity capital $(1 - \delta)K$.) Taking the present worth relative to the owner's discount rate k' gives

$$PW_{k'}\{S_C - S_E\} = (1 - \delta)(\epsilon - k')PW_{k'}\{X\}PW_{k'}\{F\}. \tag{14}$$

From this we can read the intuitively plausible result that the owner prefers capitalization over expensing if and only if his discount rate k' is less than the rate of return on equity ϵ.

It is interesting to contrast the owner of a regulated firm, considered above, with the owner of an unregulated firm. For the unregulated firm we take the revenue R as given, rather than the rate of return ρ. Then equation (14) is replaced by

$$PW_{k'}\{S_C - S_E\} = (\delta - \tau)\left(k' - \frac{(1-\tau)\delta i}{\delta - \tau}\right) PW_{k'}\{X\} PW_{k'}\{F\}. \quad (14a)$$

The present worth can be positive, i.e., capitalization can be preferable to the owner, only if

$$\delta > \tau,$$

i.e., only if the debt ratio is higher than the tax rate. Under present circumstances this would rarely be the case for an unregulated firm. Thus we would expect unregulated owners generally to favor expensing and to capitalize only as much as necessary to remain solvent.

A parenthetical comment: to understand equation (14a) intuitively, it is useful to decompose the "loan" into the sum

$$\delta(\dot{X} - iX) - \tau(\dot{X} - \delta iX).$$

The first term is a loan *to* the owner at interest rate i; the second is a loan *from* the owner at the lower interest rate δi. The first term is a debt effect; the second is a tax effect.

☐ **Tax collector.** The tax collector is interested in the effect on taxes, which (see the Appendix) is

$$T_C - T_E = (a - \rho)X_E{}^*F = \frac{\tau}{1-\tau}(\rho - i\delta)X_E{}^*F. \quad (15)$$

Taxes are always greater under capitalization. Incidentally, the dependence of tax on the depreciation rate is clearly shown by this formula. The faster the depreciation, the smaller the net plant X at any time, hence the lower the tax. This assumes, of course, that the same depreciation schedule is used for book and tax purposes.

A numerical example is displayed in Figure 3. The differentials $R_C - R_E$, $\dot{K}_{OC} - \dot{K}_{OE}$, $T_C - T_E$, and $S_C - S_E$ are graphed as functions of time for an exponentially increasing $F(t)$. This case would occur, for example, if an exponentially growing regulated firm decided at time $t = 0$ to transfer a fixed fraction of its expenditures from expense accounts to capital accounts.

5. Book and tax depreciation difference

■ Present law permits public utilities to use an accelerated depreciation schedule for tax purposes, while retaining an unaccelerated schedule for book and regulatory purposes. This results in lower taxes early in the life of a capitalized item and higher taxes later. The tax differential is treated in either of two ways: it is allowed to "flow through" to revenues or it is set aside in a "normalization" reserve, which is used for capital but is not included in the rate base.[7] The legal and other merits of the two treatments are currently a matter of considerable controversy.

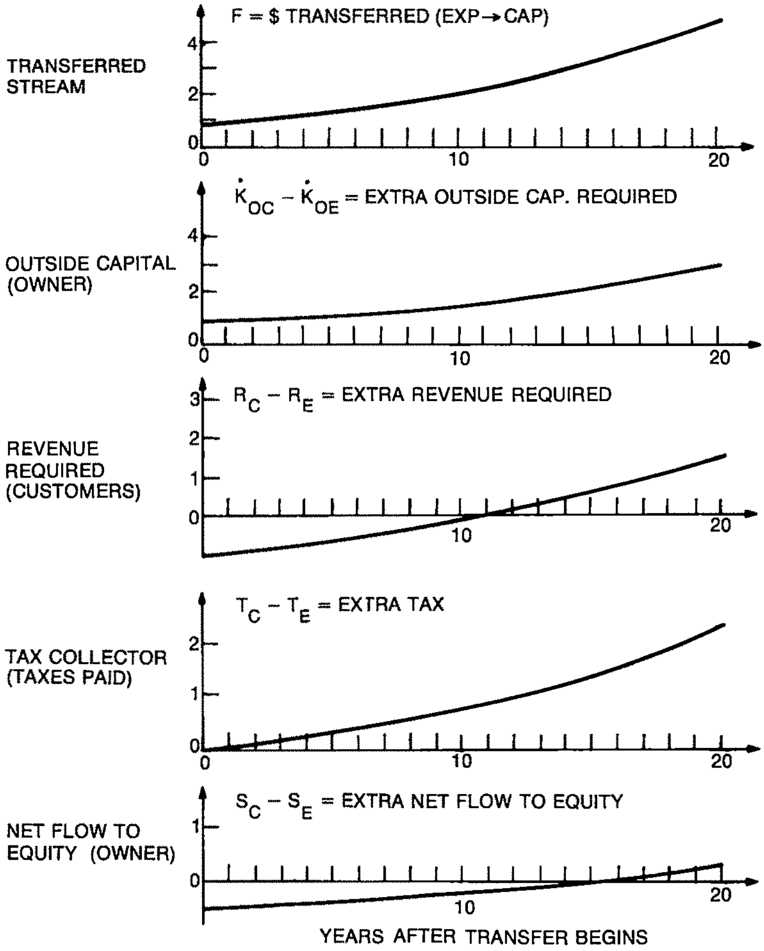
To take account of these possibilities our results must be modified as follows. (The subscripts B and T refer to book and tax depreciation.)

---

[7] See [5] for the tax law. See Linhart [7] for a discussion of alternative methods of accounting for accelerated tax depreciation.

FIGURE 3

EFFECTS OF TRANSFERRING STREAM F(t) FROM EXPENSE TO CAPITALIZATION IN THIS EXAMPLE, F(t) IS EXPONENTIAL THIS WOULD BE THE CASE, FOR EXAMPLE, IF AN EXPONENTIALLY GROWING FIRM TRANSFERRED A FIXED FRACTION OF ITS EXPENDITURES ASSUMPTIONS. REGULATED FIRM WITH $F = e^{0.08t}$, $\rho = 0.09$, $\delta = 0.45$, $i = 0.06$, $\tau = 0.50$, $d = 0.06$, USING STRAIGHT LINE DEPRECIATION WITH 20 YEAR LIFE



Accelerated tax depreciation with flow-through:

Revenues:

$$R_C - R_E = \qquad (16)$$

$$\left[\left(\frac{\rho - \tau i\delta}{1-\tau}\right)X_B - \dot{X}_B\right] *F + \frac{\tau}{1-\tau}(\dot{X}_T - \dot{X}_B)*F$$

Outside capital:

$$\dot{K}_{OC} - \dot{K}_{OE} = -(rX_B - \dot{X}_B)*F \qquad (17)$$

Taxes:

$$T_C - T_E = \frac{\tau}{1-\tau}((\rho - i\delta)X_B*F + (\dot{X}_T - \dot{X}_B)*F). \qquad (18)$$

The revenue differential is the same as before except for an added term. [See equations (8) and (9).[8]] This term has the effect of augmenting both the near-term revenue saving and the far-term revenue penalty resulting from capitalization. Note that the differential in outside capital requirements is unaffected. [Compare equations (17) and (13).]

Accelerated tax depreciation with normalization:

Revenues:

$$R_C - R_B =$$

$$\left[\left(\frac{\rho - \tau i\delta}{1 - \tau}\right)X_B - \dot{X}_B\right]*F + \frac{1 - \tau}{\tau}(\rho - \delta i\tau)(X_T - X_B)*F. \tag{19}$$

Outside capital:

$$\dot{K}_{OC} - \dot{K}_{OB} = -(rX_B - \dot{X}_B)*F + \tau[(\dot{X}_T - \dot{X}_B) - r(\dot{X}_T - \dot{X}_B)]*F. \tag{20}$$

Taxes:

$$T_C - T_B = \frac{\tau}{1 - \tau}(\rho - i\delta)X_B*F - \tau(X_B - X_T)*F + \tau(\dot{X}_T - \dot{X}_B)*F. \tag{21}$$

If $\rho > \tau i\delta$, which is true in any reasonable case, the added term in equation (19) is negative for all time. Also the added term in equation (20) is negative if r is relatively small and F grows relatively rapidly.

In general, accelerated depreciation favors capitalization. The benefit to revenue, however, is different in magnitude and time distribution for flow-through and normalization. Normalization relieves outside capital requirements; flow-through does not.

Appendix

■ In this Appendix we define some properties of the convolution, and use them to derive various formulae and results employed in the text.

☐ Convolution.[9]

Definition: $A*B \equiv \int_0^t A(t - t')B(t')dt'.$

The following properties are easily verified:

$A*B = B*A$

$\hat{\delta}*A = A,\ \hat{\delta} =$ Dirac delta function[10]

---

[8] In equation (16) we cannot substitute the pre-tax rate of return a for the expression $(\rho - \tau i\delta)/(1 - \tau)$ because a is no longer equal to this, but involves convolutions with X, etc.

[9] See Widder [9].

[10] $\hat{\delta}(t)$ is a unit impulse at $t = 0$. It is the derivative of the unit step function. If $s(t) = 0$ for $t < 0$ and $s(t) = 1$ for $t > 0$, then $\hat{\delta}(t) = \dot{s}(t)$.

$$
\begin{aligned}
&1*A = \int_0^t A\,dt' \\
&(A*B)^{\cdot} = \dot{A}*B \text{ if } A(0) = 0 \\
&PW_k(A*B) = PW_k(A)\cdot PW_k(B) \\
&PW_k(\dot{A}) = A(0) + kPW_k(A).
\end{aligned}
$$

☐ **Depreciation and capital.** From the definition of $X_B(t)$ [equation (6)]:

$$\dot{X}_B(t) = \hat{\delta}(t) - d_B(t).$$

By the convolution properties and equation (5a):

$$\dot{X}_B*C = \hat{\delta}*C - d_B*C = C - D_B. \tag{A1}$$

Since $\dot{K} = \dot{K}_O + \dot{K}_I$ (Figure 1), equation (2) yields

$$C = B_B + \dot{K}. \tag{A2}$$

Combining (A1) and (A2) gives

$$\dot{K} = \dot{X}_B*C. \tag{A3}$$

Assuming capital is zero at $t = 0$, the convolution properties also give

$$K = X_B*C. \tag{A4}$$

This says that capital equals the sum of net plant over all vintages.

☐ **Basic forms.** We shall take E, C, X_B, X_T, ρ, i, d, δ, τ as given functions of time (X_B and X_T stand for unit net plant under book and tax depreciation respectively), and calculate other quantities in terms of them.

The pre-tax rate of return a: From the definitions of ρ and a we get

$$aK = \rho K + T.$$

Substituting for T by (4) and then for $R - E$ by (1) gives

$$aK = \rho K + \tau(aK + D_B - D_T - I).$$

Solving for a, and using (A1), (A4), and the definition of i, we obtain

$$a = \frac{\rho - \tau i\delta}{1 - \tau} + \frac{\tau}{1 - \tau}\left[\frac{(\dot{X}_T - \dot{X}_B)*C}{X_B*C}\right]. \tag{A5}$$

For book and tax depreciation the same, the second term is zero. This gives equation (8).

The revenue R: Equation (1) can be written

$$R = E + aK + D_B.$$

Substituting for a, K, and D_B by means of (A5), (A4), and (1) gives

$$
\begin{aligned}
R = E + C + \left[\frac{\rho - \tau i\delta}{1 - \tau} X_B - \dot{X}_B\right]*C \qquad\qquad (A6)\\
+ \frac{\tau}{1 - \tau}(\dot{X}_T - \dot{X}_B)*C.
\end{aligned}
$$

For book and tax depreciation the same, the last term is zero and a can be substituted for the expression $(\rho - \tau i\delta)/(1 - \tau)$. Subtracting F from E and adding F to C gives equation (9).

Outside capital requirements. Using $\dot{K} = \dot{K}_O + \dot{K}_I$, we can rewrite equation (3) as

$$\dot{K}_O = \dot{K} - A + T + I + D.$$

Appropriate substitutions (definitions of ρ, i, d) give

$$\dot{K}_O = -[(\rho - i\delta - d(1 - \delta))X_B - \dot{X}_B]{*}C. \qquad \text{(A7)}$$

Tax. The tax formula is conveniently derived by rewriting the definition of ρ as

$$T = A - \rho K = (a - \rho)K,$$

and substituting (A5) and (A4):

$$T = \frac{\tau}{1 - \tau}[(\rho - i\delta)X_B{*}C + (\dot{X}_T - \dot{X}_B){*}C]. \qquad \text{(A8)}$$

Net flow to equity. Using the definitions of $\dot{K}_O$ and ρ,

$$\dot{K}_O = \dot{K} - \rho K + I + D.$$

Substituting this into

$$S = D - (\dot{K}_O - \delta\dot{K}),$$

and then using $K = X{*}C$, etc., yields

$$S = -[(1 - \delta)\dot{X} - (\rho - i\delta)X]{*}C.$$

For the unregulated firm (R given instead of ρ) we use equation (1) in place of the definition of ρ and obtain

$$S = (1 - \tau)(R - E - P) + [(\delta - \tau)\dot{X} - (1 - \tau)\delta i X]{*}C,$$

from which the equation stated in the text follows.

☐ **Normalization.** To handle this case, the model must be modified to take account of the normalization reserve. If $\dot{N}$ is the flow to the normalization reserve, then

$$\dot{N} = T_B - T_T = \tau(D_T - D_B).$$

Equations (2) and (3) and the definition of ρ must be replaced by:

$$\begin{aligned} C &= D_B + \dot{K}_O + \dot{K}_I + \dot{N} \\ \dot{K}_I &= A - I - T - D - \dot{N} \\ \rho &= (A - T - \dot{N})/K \end{aligned}$$

Substitutions similar to those outlined above yield the results stated in the text.

References

1. Averch, H. and Johnson, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.

2. Bailey, E. E. and Malone, J. C. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129–142.
3. Cootner, P. H. and Holland, D. M. "Rate of Return and Business Risk." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 211–226.
4. Davis, E. G. "A Dynamic Model of the Regulated Firm with a Price Adjustment Mechanism." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 270–282.
5. "Internal Revenue Code, 1954." United States Code, Title 26, Section 167.
6. Klevorick, A. K. "The Behavior of a Firm Subject to Stochastic Regulatory Review." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 57–88.
7. Linhart, P. B. "Some Analytical Results on Tax Depreciation." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 82–112.
8. Sharpe, W. F. *Portfolio Theory and Capital Markets.* New York: McGraw-Hill, 1970.
9. Widder, D. V. *The Laplace Transform.* Princeton: Princeton Univ. Press, 1941.

Alternative investment models for firms in the electric utilities industry

Stephen C. Peck

Assistant Professor of Economics
University of California, Berkeley

In this paper an analysis is made of the investments in turbogenerator sets made by a sample of 15 firms in the electric utilities industry for the period 1948 through 1969. Two models of firm investment are proposed and tested. The first is directly related to an earlier work of Chenery, in which he pointed out that the existence of economies of scale would lead to lumpy investments. This model is consistent with the individual firm data. The second model is the familiar distributed lag model which predicts that investment will take place smoothly. This model is inconsistent with the individual firm data. In addition, predictions of the aggregate investment of the 15 firms are made with the lumpy model and these fit about as well as a distributed lag model estimated with the aggregate data. Bayesian methods of estimation and inference are used throughout to arrive at these conclusions.

1. Introduction

■ In 1952 Chenery wrote a classic paper in which he asserted that the major explanations of investment behavior ignored economies of scale and the indivisibilities on which they are based.[1] In one sense, this statement is no longer true. Normative models involving economies of scale and lumpiness of investment have been formulated by Manne and by Massé and Gibrat[2] among others. Manne used his model for planning plant size and location for various industries in India, and Massé and Gibrat used their model for planning investments in the French electricity industry. But, in a very important sense, his statement is as true today as it was in 1952. There have been no studies of the investments made by individual firms, where a model explaining lumpy investments was

---

Stephen C. Peck received the B.A. from Cambridge University (1965), the M.Sc. (Econ.) from the London School of Economics (1967), and the M.B.A. and Ph.D. from the University of Chicago (1968 and 1973, respectively). Currently he is continuing research in the area of this article and is studying tender offers and mergers.

This paper is based on my thesis and was partially financed by NSF Grant GS-2347. I am indebted to Professors A. Zellner, J. P. Gould, P. Pashigian, M. Miller, C. R. Nelson, M. Nerlove, and V. Zarnowitz of the University of Chicago, Professors D. Gaskins and T. Keeler of the University of California, Berkeley, and to a referee. The Institute of Business and Economic Research, University of California, Berkeley, provided secretarial assistance.

[1] In [4].

[2] In [8] and [20].

formulated, estimated, and used for purposes of prediction. This paper is an attempt to remedy this lack.

An analysis is made of the investment in turbogenerator sets by a sample of fifteen firms in the electric utilities industry during the period 1947 through 1968. Two alternative models are proposed for the analysis of these investments. The first, which takes account of economies of scale is a modification of a model proposed by d'Aspremont, Gabszewicz, and Vial.[3] The other is closely related to the distributed lag model of Jorgenson.[4]

A Bayesian approach to estimation and inference is used in this paper as opposed to the more familiar classical approach. To a Bayesian statistician the parameters of a model are random variables and it is natural to characterize beliefs about the magnitudes of parameters by probability distributions. His task is to demonstrate how the data modified the beliefs he held about the parameters' magnitudes. The Bayesian starts off his analysis with a prior joint probability distribution of the parameters of the model which expresses his beliefs before the data are observed; he then derives the conditional probability distribution of the data given the parameters and the likelihood function, which encapsulates all the relevant information in the data. Then he combines the two by using Bayes' Theorem to yield the posterior joint probability distribution of the parameters. The posterior joint distribution thus includes all the relevant information about the parameters which could be gleaned from the investigator's prior beliefs and the data. Normally the investigator will be interested in the posterior probability distribution of only a subset of the model's parameters. The uninteresting parameters can be integrated out to yield the posterior distribution of interest. If the integration cannot be performed analytically, it can usually be carried out numerically. Thus the Bayesian approach provides a convenient resolution of the problem of nuisance parameters.

One feature of Bayesian data analysis which deemphasizes its conflict with classical analysis is that in certain frequently encountered situations the mean of the posterior distribution of a parameter is numerically equal to the ordinary least squares estimate of that parameter. An important example occurs in the analysis of the univariate regression model with normal error terms. The parameters of this model are the regression coefficients and the variance of the error term. If the Bayesian starts with a noninformative joint prior density function for the parameters, then the marginal posterior density function of any regression coefficient is of the Student-t form with a mean at the ordinary least squares estimate. In this paper there are a number of ordinary least squares estimations; their results should be given a Bayesian interpretation so as to be consistent with the other explicitly Bayesian analyses reported.

The paper has been organized as follows. Section 2 describes the data sample and the criteria used for the choice of the constituent firms. Section 3 introduces a theory of lumpy investment due to d'Aspremont *et al.* The model is modified in Section 4 to allow for variable costs of production and embodied technical change. Section 5 introduces the distributed lag model, the major competing theory of investment behavior in this paper. It is shown that the model collapses to a fixed lag model when dealing with the lumpy investments of individual firms. In order to estimate the modified d'Aspremont model using individual firm data, it is necessary that the firms act independently; Section 6 discusses some tests which were made to determine whether this was so. The modified d'Aspremont model was estimated in three steps. These are described in Sections 7, 8, and 9. The growth of demand for each of the fifteen firms is investigated in Section 7. In Section 8 the posterior distribution of the rate of depreciation is derived from an equation relating the intervals between successive installations and the size of the installed turbogenerators. The posterior distributions of the economies of scale parameter and of the discount rate are derived in Section 9 from an equation relating the intervals between installations and the firms' growth rates. In this section an identification problem is resolved by introducing prior information concerning the economies of scale parameter taken from a published study of the production function of electricity. There is also a discussion of the reasonableness of the posterior distributions. In Section 10 various predictions of investment for each firm are made with the modified d'Aspremont model and are added to form predicted annual aggregate investment series. These are then compared with the actual series. Sections 11 and 12 report the empirical analysis of the more familiar investment model. Section 11 shows how the fixed lag model of individual firm investment is refuted by the data. Section 12 reports the estimation of a distributed lag model of aggregate investment. A comparison of the empirical results of the two models is made in Section 13. Finally Section 14 reports conclusions and discusses implications for further work.

---

[3] In [5].

[4] In [11].

2. The data sample

■ A typical electric utility has a number of plants scattered around its service area. At each plant there are usually a number of turbogenerator sets not necessarily installed at the same time. A turbogenerator set is the turbogenerator and its boiler. Turbogenerator sets are an important item in the capital stock of a utility. For the fifteen firms in the sample, the ratio "value of turbogenerator units in service plus boiler plant equipment to the value of total electric utility plant in service," as of December 31, 1957, ranged from 20 percent to 38 percent with the average being around 27 percent.

The following three criteria were used in choosing the firms in the sample:

(1) Firms were chosen only if their plants were close together. In this way multiplant firms could be treated as if all their turbogenerators were in one central location and were used to supply the whole market. The approximate radius of the smallest circle that would enclose all the plants of each firm in the sample is 22.5 miles. The approximate radius of the smallest circle that would enclose all the plants of all but three of the firms in the sample is 12.5 miles.

(2) The sizes of the firms in the sample covered as wide a range as possible subject to criterion (1). The firms and their sizes on December 31, 1964, are shown in Table 1. Dallas Power and Light Company was the largest firm in the sample and was approximately the thirtieth largest firm in the industry in 1964. Madison Gas and Electric Company was the smallest firm in the sample and was smaller than the hundredth largest firm in the industry in 1964.[5]

TABLE 1

SAMPLE OF FIFTEEN ELECTRIC UTILITIES RANKED BY CAPACITY AS OF DECEMBER 31, 1964

FIRM	CAPACITY IN MEGAWATTS*
MADISON GAS AND ELECTRIC COMPANY	209 0
SAVANNAH ELECTRIC AND POWER COMPANY	224 4
EL PASO ELECTRIC COMPANY	398 2
CENTRAL ILLINOIS LIGHT COMPANY	487 5
NEW ORLEANS PUBLIC SERVICE, INC	783 5
INDIANAPOLIS POWER AND LIGHT COMPANY	832 3
DAYTON POWER AND LIGHT COMPANY	861 1
TAMPA ELECTRIC COMPANY	913 6
LOUISVILLE GAS AND ELECTRIC COMPANY	1,043.1
SAN DIEGO GAS AND ELECTRIC COMPANY	1,148 1
BOSTON EDISON COMPANY	1,314 4
DUQUESNE LIGHT COMPANY	1,445 5
CINCINNATI GAS AND ELECTRIC COMPANY	1,498.9
BALTIMORE GAS AND ELECTRIC COMPANY	1,507 2
DALLAS POWER AND LIGHT COMPANY	1,707 2

SOURCE FEDERAL POWER COMMISSION, *STATISTICS OF ELECTRIC UTILITIES IN UNITED STATES, CLASSES A AND B PRIVATELY OWNED COMPANIES* (WASHINGTON GOVERNMENT PRINTING OFFICE, 1965).

* THE SIZE OF THE FIRM IS MEASURED BY THE TOTAL PRODUCTIVE CAPACITY OWNED BY THE FIRM.

(3) No firm was included that had made a substantial investment in hydroelectric or atomic capacity during the period 1947 through 1966.

Many of the firms in the sample faced rapidly increasing demands in the postwar period. If the period is split into two subperiods, 1947 through 1957 and 1958 through 1968, the growth of peak load is well described as piecewise exponential. The estimated annual rates of growth vary from 0.0409 to 0.1445.

The investments in turbogenerator sets undertaken by any firm took place at discrete and often widely dispersed points of time. Also the investments of each firm followed a particular pattern. They tended to increase in size with time. At any given time, however, there was a wide range in the size of the investments undertaken by different firms. In Table 2, the investments made by the firms are displayed.

3. A theory of lumpy investment

■ All of the firms in the sample were subject to regulation, sometimes by State Power Commissions and sometimes by the munici-

---

[5] This information was inferred from Table 3-1 of W. R. Hughes, "Scale Frontiers in Electric Power," in [3].

TABLE 2

INVESTMENTS OF FIFTEEN ELECTRIC UTILITIES, 1947-1969

YEAR	MADISON	SAVANNAH	EL PASO	CENTRAL ILLINOIS	NEW ORLEANS	INDIANAPOLIS	DAYTON
1947					1×40 0 JULY 1a IND CNL	1×35 0 JULY 1a STOUT	
1948	1×20 0 JULY 1a BLOUNT				1×43 0 JULY 1a IND CNL.		1×60 0 JULY 15b HUTCHINGS
1949		1×20 0 NOV 15b RIVERSIDE	1×20 0 MAY 15 RIO GRANDE	1×35.0 JULY 1 WALLACE		1×40 0 FEB 23 WHITE RIVER	1×60.0 MAR 15b HUTCHINGS
1950					1×40.0 APR 5 IND CNL.	1×40 0 MAY 15b WHITE RIVER	1×60 0 DEC 15b HUTCHINGS
1951			1×30.0 APR 15b RIO GRANDE			1×40 0 NOV. 26 WHITE RIVER	1×60 0 FEB 15b HUTCHINGS 1×30 0 DEC. 28 TAIT
1952				1×60 0 JULY 18 WALLACE	1×25 0 SEPT 15b MKT. ST.		1×60 0 NOV. 19 HUTCHINGS
1953	1×30 0 SEPT 23 BLOUNT					1×60 0 JAN. 15b WHITE RIVER 1×60.0 DEC 15b WHITE RIVER	1×60 0 SEPT 15b HUTCHINGS
1954		1×20 0 JUNE 15b RIVERSIDE	1×30.0 NOV 15b RIO GRANDE		1×81 3 NOV 15 IND. CNL		
1955							
1956		1×30 0 JULY 15b RIVERSIDE				1×100 0 OCT 15b WHITE RIVER	
1957	1×44 0 JUNE 15b BLOUNT		1×40 0 JUNE 15b RIO GRANDE		1×100 0 APR 18 MICHOUD		
1958		1×44 0 JUNE 15b WENTWORTH	1×40 0 JUNE 15b RIO GRANDE	1×100 0 APR. 1 WALLACE		1×100 0 JUNE 15b STOUT	1×130 0 AUG. 15b TAIT
1959							1×130 0 JUNE 2 TAIT
1960			1×75 0 MAY 15b NEWMAN	1×125 0 MAY 1 EDWARDS			
1961	1×44 0 AUG 15b BLOUNT	1×46 0 DEC 15b WENTWORTH				1×100 0 MAY 15b STOUT	
1962							

TABLE 2 CONTINUED

TAMPA	LOUISVILLE	SAN DIEGO	BOSTON EDISON	DUQUESNE	CINCINNATI	BALTIMORE	DALLAS
	1×60 0 JULY 1[a] PADDY'S RUN						
1×30 0 JULY 15[b] HOOKERS PT		1×50.0 JULY 1[a] SILV GATE			1×35 0 (TOPPING) JULY 1[a] WEST END	1×60 0 AUG. 15[b] RIVERSIDE	
	1×60 0 DEC 1 PADDY'S RUN		1×81 25 OCT 1 EDGAR	1×65 0 OCT 18 PHILLIPS	1×80 0 NOV 25 COLUMBIA		1×60 0 MAY 15[b] MTN CRK
1×30 0 JAN 15[c] HOOKERS PT	1×60.0 JULY 15[b] PADDY'S RUN	1×50 0 JULY 1 SILV GATE		1×65 0 SEPT 15[b] PHILLIPS		1×66 0 OCT 15[b] WESTPORT	1×25 0 MAY 15[b] MTN CRK.
1×30 0 DEC 15[c] HOOKERS PT							
						1×75 0 SEPT. 15[b] RIVERSIDE	1×60 0 FEB 15[b] DALLAS
	1×65 0 DEC. 14 PADDY'S RUN	1×50.0 SEPT 15[b] SILV GATE	1×81 25 AUG 1 EDGAR	1×80 0 JUNE 1 ELRAMA	1×100.0 MAY 1 BECKJORD	1×100 0 DEC. 15 GOULD	
1×40 0 OCT 15[b] HOOKERS PT				1×80 0 MAR 15[b] ELRAMA	1×100.0 JULY 15[b] BECKJORD	1×75 0 SEPT. 15[b] RIVERSIDE	1×65 0 JUNE 19 PARKDALE
	1×90.0 NOV. 1 CANE RUN	1×100.0 NOV. 23 ENCINA	1×81 25 SEPT. 1 EDGAR	1×100 0 SEPT 1 ELRAMA	1×125 0 DEC. 15 BECKJORD		1×60.0 MAY 15[b] DALLAS
1×66 0 MAY 15[b] HOOKERS PT.							1×100 0 MAR 31 PARKDALE
	1×100.0 JAN. 24 CANE RUN	1×100.0 JULY 26 ENCINA		1×156 0 FEB. 29 PHILLIPS		1×125 0 FEB. 15 WAGNER	1×125 0 FEB. 15[b] MTN CRK.
1×120 0 AUG 15[b] GANNON			1×125.0 DEC 4 MYSTIC	1×100.0[d] DEC 15 SHIPPINGPORT			1×108.8 MAR 15 PARKDALE
1×120 0 NOV 15[b] GANNON	1×125 0 JULY 15 CANE RUN	1×100 0 AUG 15[b] ENCINA			1×165 0 JULY 15[b] BECKJORD		1×108.8 MAR 15[b] MTN CRK
			1×125.0 OCT. 15[b] MYSTIC			1×134.0 JAN 21 WAGNER	1×176 8 DEC. 15[b] NO. LAKE
1×161.0 OCT 15[b] GANNON		1×136 5 JULY 23 SO BAY		1×165 0 NOV. 1 ELRAMA	1×162 3 NOV 15[b] COLUMBIA		
			1×156 3 MAY 1 MYSTIC			1×190.4 JULY 18 CRANE	1×170 5 DEC 15[b] NO. LAKE
	1×163 2 MAY 15[b] CANE RUN	1×136.5 JUNE 15[b] SO BAY			1×244 8 NOV 15[b] BECKJORD		

TABLE 2 CONTINUED

YEAR	MADISON	SAVANNAH	EL PASO	CENTRAL ILLINOIS	NEW ORLEANS	INDIANAPOLIS	DAYTON
1963			1×81 6 MAY 15[b] NEWMAN		1×261 8 FEB 3 MICHOUD		
1964							
1965		1×103 5 JUNE 15[b] WENTWORTH					
1966			1×122 0 MAR 15[b] NEWMAN				
1967					1×582 3 AUG. 15[b] MICHOUD	1×253.4 JUNE 15[b] PETERSBURG	
1968				1×280 0 JULY 15[b] EDWARDS			
1969							

SOURCE PRIMARY DATA ARE FROM U S. FEDERAL POWER COMMISSION, *STEAM ELECTRIC PLANTS: CONSTRUCTION COST AND ANNUAL PRODUCTION EXPENSES* (WASHINGTON, D.C GOVERNMENT PRINTING OFFICE, 1948-1970) SECONDARY SOURCES ARE THE COMPANIES' ANNUAL REPORTS AND *MOODY'S PUBLIC UTILITY MANUAL* (NEW YORK MOODY'S INVESTORS SERVICE, 1946-1969).

THE SIZE, DATE, AND PLANT LOCATION OF EACH INVESTMENT ARE PRESENTED SINCE THEY ARE USED ONLY FOR PEAKING PURPOSES, GAS TURBINE UNITS ARE NOT TABULATED THE SIZE OF A TURBOGENERATOR SET IS MEASURED BY ITS RATED CAPACITY AT THE TIME THE INVESTMENT WAS MADE (CAUTION SHOULD BE EXERCISED IN DRAWING FINE INFERENCES FROM THE SIZES OF THE TURBOGENERATORS NAMEPLATE CAPABILITY IS AN IMPERFECT MEASURE OF THE TRUE CAPACITY OF A TURBOGENERATOR) THE NOTATION 1×100 0 MEANS THAT THE FIRM INSTALLED ONE WHOLLY OWNED TURBOGENERATOR OF RATED CAPACITY 100 0 Mw TOWARD THE END OF THE SAMPLE PERIOD, SUCH ENTRIES AS 0 2×820 0 FOR BALTIMORE GAS AND ELECTRIC CO IMPLY THAT BALTIMORE WAS THE CO-OWNER OF A TURBOGENERATOR OF CAPACITY 820 0 Mw, AND THAT ITS SHARE OF THE CAPACITY OF THAT TURBOGENERATOR WAS 20 PERCENT.

[a] BECAUSE ONLY THE YEAR OF INSTALLATION WAS AVAILABLE, THE DATE OF INSTALLATION WAS SET TO THE MIDDLE OF THE YEAR—JULY 1

[b] ONLY THE MONTH OF INSTALLATION WAS AVAILABLE, SO THE DATE OF INSTALLATION WAS SET TO THE MIDDLE OF THE MONTH

[c] AN APPROXIMATE DATE WAS ASSIGNED BECAUSE INVESTMENT INFORMATION GAVE ONLY AN APPROXIMATE TIME, SUCH AS THE FALL OF A PARTICULAR YEAR

[d] ATOMIC TURBOGENERATOR SET.

palities which they served. It was assumed that the typical utility was regulated in such a manner that it had to supply the exogenously determined demand for electricity within its service area at minimum cost. Given this assumption, a model which seemed appropriate for the analysis of investments in turbogenerator sets is that developed by d'Aspremont, Gabszewicz, and Vial.[6] Their model can be regarded as an extension of a model of Srinivasan,[7] which is an extension of certain elements of Manne's model,[8] itself an extension of certain aspects of Chenery's work.[9]

The following assumptions are made:

(1) Demand, D_t, grows geometrically at a rate g.

---

[6] In [5].

[7] In [27].

[8] In [18].

[9] In [4].

TABLE 2 CONTINUED

TAMPA	LOUISVILLE	SAN DIEGO	BOSTON EDISON	DUQUESNE	CINCINNATI	BALTIMORE	DALLAS
						1×209 5 FEB 15b CRANE	
1×187 5 JULY 1a GANNON		1×210 5 SEPT 15b SO BAY					1×361 3 DEC 15b NO LAKE
1×239 4 OCT 15b GANNON			1×359 0 AUG 15 NEW BOSTON				
	1×209 4 JUNE 15b CANE RUN					1×359 0 AUG 15b WAGNER	
1×414 0 NOV 15b GANNON		0 2×450 0 (D) ONOFRE	1×359 0 SEPT 1c NEW BOSTON	½×576 0d JULY 1a FT MARTIN		2×820 0 AUG 24 KEYSTONE-SHELOCTA	1×580 5 MAR 15b MTN CRK
						.2×820.0 JULY 23 KEYSTONE-SHELOCTA	
					37×430.0 JULY 15b BECKJORD		

(2) The depreciation of the capital stock of the firm is proportional to its working capacity and occurs at a rate δ.
(3) Demand is never greater than capacity, C_t.
(4) The planning horizon is infinite.
(5) The installation costs that result from a single capacity increment of size z are assumed to be given by a cost relationship of the form kz^a ($k > 0$, $0 < a < 1$).
(6) There is no technical change.
(7) There is a positive discount rate r.

Given these assumptions, the intertemporal cost minimizing strategy is to make an investment when capacity equals demand, wait until demand has grown so that it again equals capacity, make another investment, and so on.[10] The magnitude of each investment, and, hence, the time between investments, are determined by the interplay of two opposing tendencies. Since there are economies of scale, there is an incentive to build a very large plant, so the installation cost per unit of capacity is very small. On the other hand, since the rate of discount is positive, a dollar today is "worth more" than a dollar in the future, so there is an incentive to build a very small plant today.

The magnitude of each investment within the cost minimizing strategy is found as follows.

Let t_n be the time at which the nth investment of magnitude z_n occurs. Starting at time t_0, at which an investment must be made, the objective is to find a set of times $\{t_1^*, t_2^*, \ldots\}$ denoted $\{t^*\}$ for which the present value of installation costs is minimized.

---

[10] The time of an investment will sometimes be referred to as a regeneration point.

The present value of installation costs for any set of installation times is

$$C\{t\} = \sum_{n=0}^{\infty} e^{-r(t_n - t_0)}\, k\, z_n{}^a. \qquad (1)$$

Now each investment has to be sufficiently large to cover the increment of demand and the depreciation of the firm's capacity from the time of this investment to the next:

$$z_n = \{D_{t_{n+1}} - D_{t_n}\} + \int_{t_n}^{t_{n+1}} \delta\, D_t\, dt$$

$$z_n = \{D_{t_0}[e^{g(t_{n+1} - t_0)} - e^{g(t_n - t_0)}]\}$$

$$+ \delta\, D_{t_0} \int_{t_n}^{t_{n+1}} e^{g(t - t_0)}\, dt$$

$$z_n = D_{t_0}\, e^{g(t_n - t_0)}\, (1 + \delta/g)[e^{g(t_{n+1} - t_n)} - 1]. \qquad (2)$$

Equation (2) may be substituted into equation (1) to yield

$$C\{t\} =$$

$$k\, D_{t_0}{}^a\, (1 + \delta/g)^a \sum_{n=0}^{\infty} e^{-(r - ag)(t_n - t_0)}\, [e^{g(t_{n+1} - t_n)} - 1]^a. \qquad (3)$$

Srinivasan[11] has proved that the cost minimizing solution of this problem possesses the property that the time between the first and second investments equals the time between the second and third investments and so on; this is referred to as a constant cycle time property. His proof rests on the recognition of two facts. First, the cost minimizing set of installation times does not depend on the initial level of demand D_{t_0} as is evident from expression (3). And second, the net present cost of installation at the time of the second investment is the same as that at the time of the first investment except that D_{t_0} is replaced by D_{t_1} in expression (3). Thus it must follow that the set of cost minimizing installation times measured from time t_0 is the same as the set of cost minimizing installation times measured from time $t_1{}^*$ or

$$t_1{}^* - t_0 = t_2{}^* - t_1{}^*,\ t_2{}^* - t_0$$

$$= t_3{}^* - t_1{}^*,\ t_3{}^* - t_0 = t_4{}^* - t_1{}^*, \ldots .$$

It follows from this set of equalities that the time between successive investments in the optimal program is the same. Suppose this cycle time is denoted x. Then the expression (3) becomes

$$C\{x\} = k\, D_{t_0}{}^a\, (1 + \delta/g)^a\, (e^{gx} - 1)^a$$

$$\sum_{n=0}^{\infty} e^{-n(r - ag)x}. \qquad (4)$$

So long as $(r - ag) > 0$, the infinite sum can be simplified to yield

---

[11] In [27].

$$C\{x\} = k\,D_{t_0}{}^{a}\ (1 + \delta/g)^{a}\,\frac{(e^{gx} - 1)^{a}}{1 - e^{-(r-ag)x}}.$$

The cost minimizing interval between investments x can be derived by methods of calculus. It can be shown that a minimum exists so long as the further restriction $0 < a < 1$ is imposed. The minimum is given as a solution of the following implicit equation in x:

$$ag[e^{(r-ag)x} - 1] = (r - ag)(1 - e^{-gx}). \qquad (5)$$

A reasonably exact approximate solution for x can be obtained in direct form by expanding the exponentials to their squared terms and solving for x. The resulting expression is

$$x = \frac{2(1 - a)}{g(1 - a^2) + ar}.$$

Thus the optimal solution of the d'Aspremont model involves making installations at equally spaced intervals of length x where x is given as a solution of equation (5). Furthermore, since demand increases geometrically, the magnitude of each turbogenerator increases geometrically also as can be seen by substituting in equation (2):

$$\begin{aligned} z_n &= D_{t_0}e^{ngx}(1 + \delta/g)(e^{gx} - 1) \\ &= D_{t_n}(1 + \delta/g)(e^{gx} - 1). \qquad (6) \end{aligned}$$

It may be seen from this expression that $z_{n+1/z_n} = e^{gx}$. In addition if two firms are growing at the same geometric rate, then the larger firm will build larger turbogenerators.

4. A modification of the d'Aspremont model

■ The model proposed above deals solely with installation costs to the complete neglect of variable costs. This is unsatisfactory for an analysis of investment in turbogenerator sets where the variable input fuel is quantitatively significant. Once a variable cost of production is introduced, the optimal solution becomes very hard to characterize. This is shown by the following argument.

Consider the first two investments made by a firm. Suppose that a policy involves turbogenerators of not necessarily different sizes, z_1 and z_2. It is evident that the investor has to solve a different problem when the first and second turbogenerators are installed. After the first installation, all the incremental output would be allocated to the first turbogenerator, whereas after the second installation, the investor must allocate the incremental output between the two turbogenerators.

The essence of Srinivasan's proof that the optimal policy involves a constant time between investments is that the investor has to solve the same problem at the first and second regeneration points. Once variable costs are introduced, this is not so. Therefore, the optimal policy does not necessarily involve a constant cycle time and it is impossible to simplify the minimand as was done in passing from equations (3) to (4).

Thus, it is not possible to introduce variable costs in an exact manner and retain the tractability of the original problem. It is now shown that the model of d'Aspremont *et al.* can be viewed as an approximation to a more complicated model where fuel and labor costs are introduced.

Let $\underline{w}$ and $\underline{p}$ be the column vectors of the inputs and the rentals of the factors of production in steam electric generation; let y be a measure of output and let A and B be measures of neutral technological change. Then, the production function of electricity is given by expression (7) and the corresponding long-run cost function of electricity is given by expression (8):[12]

$$y = Af(\underline{w}) \tag{7}$$

$$c(y,\underline{p}) = \{\min_{w} p'\underline{w} \text{ subject to } Af(\underline{w}) > y\} = Bg(\underline{p})y^{a}. \tag{8}$$

The production function is assumed to be homogeneous of degree $1/a$ which implies that the cost function is homogeneous of degree a.[13] The empirical support of this assumption is discussed later in Section 9.

Now suppose that the real factor rentals are anticipated to remain the same and that a turbogenerator is installed which, when operating at a rate z is on the long-run cost function; its total cost rate is given by expression (8). Suppose, also, that technical change is entirely embodied and further, that it is anticipated that the turbogenerator will operate at a rate z in perpetuity.[14]

Then the long-run discounted cost of operation of such a turbogenerator is:

$$d(z) = \int_0^\infty e^{-rt}c(z,\underline{p})dt = \int_0^\infty e^{-rt}Bg(\underline{p})z^a dt = \frac{Bg(\underline{p})z^a}{r}$$

or

$$d(z) = kz^a \quad \text{where} \quad k = \frac{Bg(\underline{p})}{r}.$$

This analysis has shown that instead of interpreting the cost kz^a as an installation cost, it may be interpreted as a discounted total cost. This makes it possible to use studies of the production function of electricity to provide prior information concerning the value of the parameter a.

Furthermore, it makes it possible to show the consequences of neutral embodied technical change. Suppose that the index B falls at a rate ψ; then the factor k will fall at the same rate:

$$k_t = k_{t_0}e^{-\psi(t-t_0)}.$$

---

[12] From the Shephard-McFadden-Uzawa duality theorem, we know that to each well-behaved production function there corresponds a cost function with certain properties. This implies that the function $g(p)$ has certain properties which we will not go into at this time.

[13] It is simple to prove that the cost function is homogeneous of degree a if and only if the production function is homogeneous of degree $1/a$.

[14] This may not be a bad approximation. New turbogenerators, once they have been checked for reliability, are operated permanently near capacity as base load units, except for regular periods of inspection. By the time they are old units and no longer used for base load, the discount factor e^{-rt} may be quite large.

The present value of installation costs will then be

$$C\{t\} = \sum_{n=0}^{\infty} e^{-r(t_n - t_0)} k_{t_0} e^{-\psi(t_n - t_0)} z_n{}^a.$$

Once the substitution for z_n has been made, this expression becomes

$$C\{t\} =$$

$$k_{t_0} D_{t_0}{}^a (1 + \delta/g)^a \sum_{n=0}^{\infty} e^{-(r+\psi-ag)(t_n - t_0)} [e^{g(t_{n+1} - t_n)} - 1]^a.$$

But this expression is the same as the minimand of the d'Aspremont model with $(r + \psi)$ replacing r. So, by analogy, the cost minimizing cycle time is given as a solution of the following equation:

$$ag[e^{(r+\psi-ag)x} - 1] = (r + \psi - ag)(1 - e^{-gx}).$$

An approximate direct solution of this equation is

$$x = \frac{2(1-a)}{g(1-a^2) + a(r+\psi)} \tag{9}$$

and the two conditions for a solution to exist are

$$r + \psi - ag > 0, \ 0 < a < 1.$$

Turbogenerator size still increases at the same rate as demand.

Much of the empirical work is concerned with equation (9). The economies of scale parameter a and the rate of technical change ψ are treated as unknown parameters whose values are the same for all the investments in the sample. The real cost of capital probably did not vary much over the period 1948 through 1968. Also, it can be shown that the cycle time is almost insensitive to changes in the growth rate of demand g.[15] It follows, therefore, that the time between each investment should be approximately equal, that larger firms should build larger turbogenerators (*ceteris paribus*), and that the size of turbogenerators installed by each firm should grow at the same rate as demand.

5. Distributed lag models of investment

■ A model of investment which has gained much acceptance in the literature is the distributed lag model. Among those who have proposed and used the model are Koyck, Chenery, Eisner, and Jorgenson.[16] Although there are different variants of the model, they all have one characteristic in common: investment responds gradually to changes in demand for the firm's output.

A variant of Jorgenson's model will be used in this paper. The major implication of such a model is that there should be an investment in almost every year. This is clearly falsified by the data.

In a situation such as this one, where capacity and output are measured in the same units (megawatts), the distributed lag model is

---

[15] a, r and ψ were assigned "reasonable" values in the implicit equation for x above. Solutions of x were obtained by numerical methods for varying values of g.

[16] In [16], [4], [7], and [11], respectively.

$$\Delta C_t = \sum_{i=0}^{\infty} \mu_i \Delta D_{t-i} + \delta C_t + u_t, \quad \mu_i \geqslant 0, \sum_{i=0}^{\infty} \mu_i = 1,$$

where ΔC_t is firm investment in period t, ΔD_{t-i} is the increment of firm output, C_t is firm capacity, and u_t is an error. The parameters of the model are δ, the physical rate of depreciation, and the set of numbers $\mu \cdot \mu_j$ is defined as the proportion of investment projects delivered j periods after they are ordered.

It is argued now that for the analysis of lumpy investments made by a firm, Jorgenson's model collapses to a fixed lag model. For we would expect that a firm which is not too large would order at most one turbogenerator at a time, and this turbogenerator either would or would not be delivered j periods later.[17] Thus the only way to apply the model as a regression model to the individual firm data is to assume that there is a fixed lag of τ periods or that

$$\mu_\tau = 1, \ \mu_i = 0 \ \ i = 0, 1, 2, \ldots i \neq \tau.$$

Thus the fixed lag model that will be used for the investigation of the behavior of the investment of individual firms is given by the following expression:

$$\Delta C_t = \Delta D_{t-\tau} + \delta C_t + u_t.$$

The major implication of such a model is that since the rate of depreciation, δ, is presumably positive and since, as will be shown, demand rises almost monotonically, each firm should make an investment in every period. If the word period is interpreted as year, this prediction of the model is obviously false, as a glance at Table 2 will show.

A distributed lag model of investment was estimated for aggregate investment of all fifteen companies partly because such a model of aggregate investment is so frequently used and partly because it provides a standard of comparison for the modified d'Aspremont model.

6. Preliminary discussion of the empirical analysis of the modified d'Aspremont model

■ The modified d'Aspremont model deals with a situation in which demand increases geometrically and investments are made to meet that demand. There is only one market and investments are made to satisfy demand in that market. In order to ensure that this was the case, the sample contained only firms whose plants were close together.

In addition it is assumed that the firm does not behave in the manner depicted in Manne's backlog case;[18] that is it does not purchase from a neighboring firm to postpone its investments in a systematic manner. According to Vennard in his book on the technical and institutional framework of the electric power industry,[19] it is common practice for neighboring firms to enter into agreements to swap electricity or sell to each other. Since this was claimed to be the case, it was necessary to check quite carefully whether imports of electricity were being used by the firms in the sample to postpone their turbogenerator installations. This question was investigated in two different and complementary ways.

---

[17] We note here that if the data had been in the form of capital expenditures on new turbogenerator sets, then it would have been meaningful to entertain a distributed lag model in which the dependent variable was investment expenditures on new turbogenerator sets. Then μ_j could have been interpreted as the proportion paid in period $t + j$ of an order made in period t.

[18] In [18].

First, an investigation was made as to whether there were large imports of energy for some time before an investment took place and none afterwards. The annual imports of energy were measured by the sum of the purchases and interchanges *in* of energy.[20] The importance of the imports was evaluated by how long they could have postponed an investment on the assumption that the firm imported electricity for exactly half the year—4,380 hours. There was a certain amount of evidence that some cycle times could have been extended due to the import of energy. In the majority of cases, the actual cycle time was not prolonged by more than 15 percent and the effect was not firm-specific; a firm would import a lot before one investment and very little before the next. There was one regularity apparent in these data, however.

A glance at Table 2 will show that some companies had abnormally long intervals between investments in the 1960s; they were Madison, Illinois, Indianapolis, Dayton, Duquesne, and Cincinnati. Except for Illinois, these abnormally long cycle times were associated with abnormally large imports of energy.

Second, the relationship between each firm's peak load and capacity was investigated to see whether peak load exceeded capacity just before each investment took place and was exceeded by capacity just after each investment.[21] This is the pattern that would be expected if the firm was systematically postponing investments by importing electricity. In fact, for one of the companies capacity was always greater than peak load, while for most of the others capacity was less than peak load for five or six years after World War II; it grew faster than peak load during some part, at least, of the postwar period and at the end of the period was greater than peak load at all times. Exceptions to this pattern were provided by those firms which had abnormally long cycle times in the 1960s. For all of these firms, except Illinois, peak load was allowed to exceed capacity in the 1960s.

---

[19] In [32].

[20] The data on purchase and interchange in were collected from FPC *Statistics of Electric Utilities, Classes A and B.* The following definition of interchange energy is to be found in U. S. Federal Power Commission, *Glossary of Important Power and Rate Terms, Abbreviations and Units of Measurement* (Washington, D. C.: U. S. Government Printing Office, 1965): "Interchange energy is electric energy received by one electric utility usually in exchange for energy delivered to another system at another time or place. Interchange energy is to be distinguished from a direct purchase or sale, although accumulated energy balances are sometimes settled in cash."

[21] The data on peak loads were gathered from Moody's *Public Utilities* and on firm capacity from FPC *Statistics of Electric Utilities, Classes A and B,* in which firm capacity is referred to as the "total productive capacity owned by the firm."

There are two conclusions of importance here. First, for most firms, capacity grew faster than peak load during some part at least of the late 1940s and early 1950s. From perusal of industry sources, it seemed that firms found it difficult to obtain turbogenerators during World War II, since much of the new output of the producing industry was devoted to military uses. Second, for most firms there was no evidence that they systematically postponed investments by importing electricity, although there was evidence that a group of firms which had abnormally long cycle times in the 1960s imported abnormally large quantities of electricity in the 1960s. So the evidence on imports does not cause a rejection of the modified d'Aspremont model.[22]

It was necessary to estimate three separate equations in the investigation of the modified d'Aspremont model. First, the rate of growth of demand of each of the fifteen firms was estimated. Then the rate of depreciation was estimated. Finally, the economies of scale parameter and the anticipated rate of technical change were estimated. For the second and third estimations, the sample of fifteen firms was split into two; subsample 1 consisted of six firms and subsample 2 consisted of the remaining nine. Extensive pretesting was carried out with subsample 1 to choose the "best" specification of the model and then the best specification was estimated with subsample 2.[23]

7. Estimating the growth of demand

■ In the pretesting, it was found that the firms seemed to respond more to demand as measured by peak load than to demand as measured by generation; that is, the model fit better with demand measured by peak load rather than by generation.[24] Accordingly, only the estimation of peak load is reported here.

To determine the annual rate of growth of peak load for each company, the following equation was estimated by the method of ordinary least squares:

$$\log D_{it} = \alpha_i + g_i(t - 1946) + u_{it} \quad i = 1, 2, \ldots 15 \qquad (10)$$
$$t = 1947, 1948, \ldots, 1969.$$

For almost every company, there was considerable serial correlation of the residuals and thus there were very low Durbin-Watson statistics. A study of the residual patterns revealed that the serial correlation was not due to cyclical fluctuations about a geometric growth path but to the choice of an inappropriate functional form. Accordingly, an arithmetic growth model was estimated by ordinary least squares:

$$D_{it} = a_i + q_i(t - 1946) + v_{it} \quad i = 1, 2, \ldots 15$$
$$t = 1947, 1948, \ldots, 1969.$$

---

[22] The details of this investigation are provided in Appendix B of my thesis [23].

[23] The firms in subsample one were Boston, Baltimore, Louisville, El Paso, Illinois, and New Orleans. The firms in subsample 2 were Dallas, Cincinnati, Duquesne, Dayton, Tampa, San Diego, Indianapolis, Savannah, and Madison.

[24] The annual peak load series were collected from Moody's *Public Utilities.* The annual total net generation series were collected from FPC *Statistics of Electric Utilities, Classes A and B.* Net generation is a measure of output which excludes the power used to run the generating stations.

Again, for almost every company there was considerable serial correlation of the residuals, and a study of the residual patterns indicated that the serial correlation was not due to cyclical fluctuations but to the choice of an inappropriate functional form.

Accordingly, the sample period was split in two and the geometric growth regression (10) was estimated for the periods, 1947 through 1958 and 1959 through 1969 for each firm. The Durbin-Watson statistics for each company and each time period were almost all within an acceptable range,[25] and it appears that for many of the firms peak load grew greatly different rates in the periods 1947 through 1958 and 1959 through 1969; frequently, the differences in growth rates were statistically significant at the 5-percent level. The ordinary least squares estimates of each firm's growth rates are shown in Table 3.

TABLE 3

ORDINARY LEAST SQUARES RESULTS FOR
PEAK LOAD 1947-1958 AND 1959-1969

FIRM	ESTIMATED GROWTH RATE		FIRM	ESTIMATED GROWTH RATE	
	1947-1958	1959-1969		1947-1958	1959-1969
DALLAS	0 1425	0 0812	SAN DIEGO	0 0886	0 0717
SAVANNAH[a]	0 0973	0 0886	NEW ORLEANS	0 0416	0 1119
CINCINNATI[a]	n a	0 0680	TAMPA	0 1445	0 0842
BOSTON	0 0409	0 0505	LOUISVILLE	0 0673	0 0712
DUQUESNE	0 0487	0 0417	INDIANAPOLIS	0 0796	0 0741
EL PASO	0 1021	0 0712	DAYTON	0 0771	0.0776
CENTRAL ILLINOIS	0 0584	0 0639	MADISON[a]	0 0833	0 0831
BALTIMORE	0 0513	0 0851			

[a] DATA ON PEAK LOADS WERE NOT REPORTED FOR SAVANNAH FOR THE YEARS 1959-1962, FOR CINCINNATI FOR THE YEARS 1947-1952 AND FOR MADISON FOR THE YEARS 1947-1950

8. Derivation of the posterior distribution of the rate of depreciation

In a previous section, it was shown that the relationship between the size of the nth turbogenerator installed, z_{in}, and the optimal cycle time between the nth and $(n+1)$th investments, x_{in}, was given by the following expression, which is a repetition of equation (6) with subscripts i added to refer to the ith firm:

$$z_{in} = D_{it_n}(1 + \delta/g_i)(eg_i^{x_{in}} - 1).$$

This expression can be solved for the optimal cycle time, x_{in}:

$$x_{in} = \frac{1}{g_i} \log\left(1 + \frac{z_{in}}{D_{it_n}(1 + \delta/g_i)}\right). \qquad (11)$$

This is the cycle time implied by the size of the turbogenerator installed and the growth of demand. It will be referred to henceforth as the anticipated cycle time to distinguish it from the actual cycle time, denoted X_{in}. Once the turbogenerator was installed, there may have been many reasons why the actual time between this investment and the next differed from the anticipated time between the two investments. It was assumed that the actual cycle time and the anticipated cycle time were related as[26]

---

[25] With only eleven or twelve observations, the power of the Durbin-Watson test was not great.

$$X_{in} = x_{in} e^{v_{in}},$$

there

$$v_{in} \sim N(0, \sigma_v^2) \quad \text{and} \quad E(v_{in} v_{jm}) = 0 \quad \text{for} \quad in \neq jm.$$

Thus the equation relating the cycle time and the rate of depreciation was

$$\log X_{in} = -\log g_i + \log\log\left(1 + \frac{z_{in}}{D_{it_n}(1 + \delta/g_i)}\right) + v_{in}.$$

The independent variables in this equation are z_{in} the size of the nth turbogenerator installed by the ith firm,[27] g_i the anticipated growth rate of peak load at the time the investment was made and D_{it_n} the predicted peak load at the time of investment. g_i was approximated by the ith firm's estimated growth rate, $\hat{g}_i$, during the period in which the investment was made.[28] D_{it_n} was approximated by $\hat{D}_{it_n}$, the predicted peak load of the ith firm at the time the investment was made. $\hat{D}_{it_n}$ was measured by

$$\hat{D}_{it_n} = \exp\left(\hat{\alpha}_i + \hat{g}_i(t_n - 1946\cdot5) + 1/2\, \hat{\sigma}_{u_i}^2\right),$$

where $\hat{\alpha}_i$, $\hat{g}_i$, and $\hat{\sigma}_{u_i}^2$ are the estimated intercept, growth rate, and error variance from the regression of the log of peak load on time during the period in which the investment was made. $\hat{g}_i$ and $\hat{D}_{it_n}$ were treated as known and error-free exogenous variables, so that errors in the variables problems were completely neglected. Thus, there is no reason to suppose that $\hat{g}_i$ and $\hat{D}_{it_n}$ are correlated with the error term v_{in}. In addition, z_{in} and v_{in} are uncorrelated, as is shown by the following argument. If z_{in} is large, then the anticipated cycle time x_{in} is long (*ceteris paribus*), but the actual cycle time X_{in} can be either longer or shorter than the anticipated cycle time . . . that is v_{in} can be either positive or negative. Thus, z_{in} and v_{in} are uncorrelated.

Bayes' theorem is used to derive the posterior distribution of the rate of depreciation. The probability density function of the ith cycle time of the nth firm is

$$p(X_{in}|\delta, \sigma_v) \propto \frac{1}{\sigma_v} \exp\left[-\frac{1}{2\sigma_v^2}\left\{\log X_{in} + \log \hat{g}_i - \log\log\left(1 + \frac{z_{in}}{\hat{D}_{it_n}(1 + \delta/\hat{g}_i)}\right)\right\}\right].$$

---

[26] In pretesting with subsample 1, it appeared that this specification of the relationship was more satisfactory than

$$X_{in} = x_{in} + v_{in}.$$

[27] The magnitude of a turbogenerator was measured in megawatts of net continuous capability when the turbogenerator was not constrained by a lack of condenser water.

[28] The assumption was made that the change in the growth rate which occurred around 1957 was unanticipated.

And the joint probability density function of all the observations, the likelihood function, is

$$p(X_{11}, X_{12}. \ldots |\delta, \sigma_v) \propto \sigma_v^{-\nu} \exp\left\{-\frac{f(\delta)}{2\sigma_v^2}\right\},$$

where

$$\nu = \sum_{i=1}^{N_f} (n_i - 1)$$

and where

$$f(\delta) = \sum_{i=1}^{N_f} \sum_{n=1}^{n_i-1} \left\{\log X_{in} + \log \hat{g}_i - \log\log\left(1 + \frac{z_{in}}{\hat{D}_{it_n}(1 + \delta/\hat{g}_i)}\right)\right\}.^2$$

All elements of the likelihood function not involving δ and σ_v are summarized in the implicit constant of proportionality.[29]

The two parameters of the model are δ and σ_v; *a priori* I was prepared to admit to no information as to their value except that $0 < \sigma_v < \infty$ and $0 < \delta < 1$. The prior distributions of δ and σ_v were assumed to be independent; the prior distribution of δ was assumed to be uniform over the interval from zero to one, and the prior distribution of log σ_v was assumed to be uniform over the interval zero to infinity. The prior joint probability density function of δ and σ_v was therefore taken to be

$$p(\delta, \sigma_v) \begin{cases} \propto \dfrac{1}{\sigma_v}, & 0 < \delta < 1 \\ = 0, & \text{otherwise.} \end{cases}$$

Bayes' theorem provides an analytical procedure for combining the prior information about the parameters with the sample information. More formally, it provides a well-defined method for transforming the prior probability density function $p(\delta, \sigma_v)$ into the posterior probability density function $p(\delta, \sigma_v|X_{11}, X_{12}\ldots)$. The statement of the theorem is

$$p(\delta, \sigma_v|X_{11}, X_{12}\ldots) \propto p(\delta, \sigma_v)p(X_{11}, X_{12}\ldots|\delta, \sigma_v).$$

Substitution yields the joint posterior density function of δ and σ_v, and analytical integration with respect to the nuisance parameter σ_v over the interval zero to infinity yields

---

[29] The explanation of the indices of the summation signs is as follows. The firm index i runs from 1 to N_f where N_f is the number of firms used to derive the likelihood function. The investment number index runs from 1 to $n_i - 1$; investment 1 refers to the first investment made by the firm after 1946 and investment $n_i - 1$ refers to the penultimate unshared investment made by firm i before 1969. The investment index number runs only to $n_i - 1$; for those firms which made no shared investment in the late 1960s, no cycle time could be computed between their last investment before 1969 and the next one, since the time of the next investment was unknown. For those firms which made a shared investment in the late 1960s, no cycle time was computed between the last unshared investment and the shared investment, because at this time the firm had begun to plan its investments jointly with some other firm or firms.

$$p(\delta|X_{11}, X_{12}\ldots) \begin{cases} \propto \dfrac{1}{[f(\delta)]^{\nu/2}}, & \text{for } 0 < \delta < 1 \\ = 0, & \text{otherwise.} \end{cases}$$

Preliminary work suggested that it was reasonable to pool observations of different firms to compute the posterior density function of δ. Details of the posterior distribution of δ together with some summary residual statistics are given in Table 4. The posterior distributions of δ for both subsamples were unimodal and symmetric. On the whole, the mean residuals for both subsamples suggested that the assumptions of independence and constant variance were satisfied.[30] The only obvious exception to this was

TABLE 4

SUMMARY STATISTICS OF THE DISTRIBUTIONS OF δ AND RESIDUAL STATISTICS

	SUBSAMPLE ONE	SUBSAMPLE TWO[b]
NUMBER OF OBSERVATIONS	42	66
DISTRIBUTION OF δ		
MEAN	0.0217	0 0340
VARIANCE	0.491×10^{-4}	0.705×10^{-4}
RESIDUAL STATISTICS FOR THE ENTIRE SUBSAMPLE[a]		
SERIAL CORRELATION COEFFICIENT	−0.07	−0.01
RESIDUAL VARIANCE	0 27	0 28

[a] THE RESIDUAL STATISTICS WERE COMPUTED FROM THE MEAN RESIDUALS IN COMPUTING THE SERIAL CORRELATION COEFFICIENT, A BREAK WAS MADE WHEN THE ADJACENT OBSERVATIONS DID NOT BELONG TO THE SAME COMPANY

[b] THE POSTERIOR DISTRIBUTION OF δ FOR FIRMS IN SUBSAMPLE TWO WAS DERIVED, WHERE THE LIKELIHOOD FUNCTION WAS SUCH THAT THE VARIANCE OF DAYTON'S CYCLE TIMES WAS NOT CONSTRAINED TO EQUAL THAT OF THE OTHER FIRMS THE MEAN AND VARIANCE OF THE P D.F WERE 0 0246 AND $0\,331 \times 10^{-4}$ AN EXAMPLE OF THIS PROCEDURE IN THE LINEAR CASE IS PROVIDED IN ZELLNER [33], pp 98-107

---

[30] Within the Bayesian framework, a residual is a random variable, since it is a function of δ, itself a random variable. In the residual analysis, the mean residual $\bar{v}_{in}$ was used, where

$$\bar{v}_{in} = E_\delta(v_{in}) = \log X_{in} + \log \hat{g}_i$$
$$- E_\delta \log\log\left(1 + \frac{z_{in}}{\hat{D}_{it_n}(1 + \delta/\hat{g}_i)}\right).$$

The last term involving the expectation with respect to δ was approximated by using the following result.

Consider a random variable δ with a mean $\bar{\delta}$ and a variance var δ. Then the mean of a function $h(\delta)$ of δ is approximated by the following expression:

$$E_\delta h(\delta) \simeq h(\bar{\delta}) + 1/2 \left.\frac{d^2h}{d\delta^2}\right|_{\delta=\bar{\delta}} \operatorname{var} \delta.$$

provided by Dayton's residuals which appeared to have larger variance than those of the other firms in subsample two.

The means of the posterior distributions are "reasonable." For instance, for subsample one the mean rate of depreciation is 2.17 percent per annum. Furthermore, the similarity of the results for subsamples one and two is striking. The difference between the rates of depreciation for the two subsamples has a mean of 0.0123 and a standard deviation of 0.0118.

No formal test was made of the null hypothesis that the rates of depreciation in the two subsamples were the same. This was not done because the posterior distribution of the difference between the two rates of depreciation was not derived. However, it is comforting to note that if the posterior distribution of each rate of depreciation were normal, it would not be possible to reject the null hypothesis at a 5-percent level of significance.

9. Derivation of the posterior distribution of the economies of scale parameter and the discount rate

■ The relationship between the optimal cycle time x and its determinants a, ψ, r, and g was shown in equation (9). If in addition it is assumed that investors may make mistakes in their choice of an optimal cycle time, the relationship between x and its determinants may be represented as[31]

$$x_{in} = \frac{2(1-a)}{g_i(1-a^2) + a(r_{in} + \psi)} e^{w_{in}},$$

where subscripts have been added to refer to the nth investment of the ith firm and where w_{in} is a normally distributed error term. The independent variables in this equation are g_i, the growth rate of peak load at the time the investment was made, and r_{in}, the firm's real cost of capital. g_i was approximated by the ith firm's estimated growth rate, $\hat{g}_i$ during the period in which the investment was made. Attempts were made to derive a real cost of capital series for the 23-year postwar period, but these proved to be unsuccessful and so it was assumed that the cost of capital was the same for all the investments of all the firms in the sample. Since this was the case, the sum of $r + \psi$ was treated in the estimation as a new parameter θ called the discount factor.

The dependent variable x_{in} was not observable; two alternative approaches were available to resolve the problem. In the first, the relationship between the actual and the anticipated cycle time could be used to derive a relationship between the actual cycle time and a, θ, and g. Since

$$X_{in} = X_{in} e^{v_{in}}.$$

It follows that

$$X_{in} = \frac{2(1-a)}{\hat{g}_i(1-a^2) + a\theta} e^{w_{in}+v_{in}}.$$

The posterior distribution of a and θ could then be derived.

The second approach is best explained if it is first assumed that the rate of depreciation δ is known. The size of the installed turbogenerator together with demand at the time of the installation and the growth rate of demand determine the anticipated cycle time as shown in equation (11). The series of anticipated cycle times could then be used as the dependent variable in deriving the posterior distribution of a and θ. In fact, only estimates of demand at the time of the investment, estimates of the growth rate in demand and the posterior distribution of the rate depreciation were available, but the second approach provided an operational way to obtain a series of approximate anticipated cycle times.

---

[31] In the pretesting it appeared that this specification of the error term was more satisfactory than one in which the error term entered additively.

The second approach was used almost exclusively in the empirical work for the following reason. Given the absence of any variation in the cost of capital, it was pointed out previously that the anticipated cycle time should be almost the same for each investment. Now the distribution of the anticipated cycle times constructed by using the turbogenerator sizes was much more peaked than the distribution of the actual cycle times.[32]

Thus the posterior joint distribution of the economies of scale parameter a and the discount factor θ was derived from the following equation:

$$\log \bar{x}_{in} = \log 2(1 - a) - \log [\hat{g}_i(1 - a^2) + a\theta] + w_{in}, \qquad (12)$$

where the dependent variable $\log \bar{x}_{in}$ was constructed according to the following formula:

$$\log \bar{x}_{in} \equiv -\log \hat{g}_i + E_\delta \log \log \left(1 + \frac{z_{in}}{\bar{D}_{it_n}(1 + \delta/\hat{g}_i)}\right).$$

Use was made of the result in note 30 to evaluate the expected value of the last expression.

Preliminary analysis with subsample one suggested that the errors w_{in} were autocorrelated; there were firm-specific effects determining the anticipated cycle times. Accordingly, the following structure was assumed for the errors

$$w_{in} = \rho w_{in-1} + \epsilon_{in}, \quad i = 1, 2 \ldots N_f, \quad n = 2, 3, \ldots n_i,$$

where

$$\epsilon_{in} \sim N(0, \sigma_\epsilon^2), \quad E(\epsilon_{in}\epsilon_{jm}) = 0 \quad in \neq jm \quad \text{and} \quad 0 < \rho < 1.$$

N_f is the number of firms in the subsample and n_i is the number of investments made by the ith firm.

---

[32] An anticipated cycle time was constructed for each investment in subsample one for which a cycle time had been computed. The means and standard deviations of the actual and anticipated cycle times were then computed. The same was done for subsample two. The results are tabulated below. They show a considerable reduction in the spread of the anticipated cycle time series, relative to the actual cycle time series.

	Subsample One		Subsample Two	
	Actual Cycle Time	Anticipated Cycle Time	Actual Cycle Time	Anticipated Cycle Time
Average	2.64	2.40	2.00	1.78
Standard Deviation	1.47	0.88	1.19	0.62

Given this specification of the errors, the probability density of the ith observation for the nth firm is

$$p(\log \bar{x}_{in}|a, \theta, \sigma_\epsilon, \rho) \propto \frac{1}{\sigma_\epsilon} \exp\left(\frac{-1}{2\sigma_\epsilon^2}\{\log \bar{x}_{in} - \log 2(1-a) + \log[\hat{g}_i(1-a^2) + a\theta] - \rho(\log \bar{x}_{in-1} - \log 2(1-a) + \log[\hat{g}_i(1-a^2) + a\theta])\}^2\right).$$

To ease the notational burden, y_{in} is defined

$$y_{in} = \log \bar{x}_{in} - \log 2(1-a) + \log[\hat{g}_i(1-a^2) + a\theta].$$

The conditional probability density function for all the investments of all the firms in the subsample, the likelihood function, is given by the following expression:[33]

$$p(\log \bar{x}_{11}, \log \bar{x}_{12} \ldots |a, \theta, \sigma_\epsilon, \rho, M_1, \ldots M_{N_f}) \propto$$

$$\frac{1}{\sigma_\epsilon^\nu} \exp\left(\frac{-1}{2\sigma_\epsilon^2}\left\{\sum_{i=1}^{N_f}(y_{i1} - M_i)^2 + \sum_{i=1}^{N_f}\sum_{n=2}^{n_i}(y_{in} - \rho y_{in-1})^2\right\}\right),$$

where

$$\nu = \sum_{i=1}^{N_f} n_i.$$

Additional preliminary work with subsample one suggested that there was an identification problem. In principle the parameters a and θ are identified in equation (12), since the smaller the value of a, the larger is the value of $(1 - a^2)$ and the more responsive the anticipated cycle time is to the value of $\hat{g}_i$. In practice the magnitude of this effect was so small that the parameters a and θ were almost not identified. To identify the parameters, some prior information was introduced concerning the magnitude of the economies of scale parameter a.

It was shown previously that the cost function kz^a could be

---

[33] The parameters M_i were introduced to deal with the first element of the autoregressive process for each firm. Thus, the probability density function of the first observation for the first firm is

$$p(\log \bar{x}_{11} \mid a, \theta, \sigma_\epsilon, \rho, y_{10}) \propto \frac{1}{\sigma_\epsilon} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_{11} - \rho y_{10})^2\right\}.$$

Now y_{10} in this expression is unobserved, and the probability density function is reparameterized by setting $\rho y_{10} = M_1$, thus

$$p(\log \bar{x}_{11} \mid a, \theta, \sigma_\epsilon, M_1) \propto \frac{1}{\sigma_\epsilon} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_{11} - M_1)^2\right\}$$

interpreted as a discounted homogeneous long-run cost function. Such a cost function could only be derived from a production function which was homogeneous of degree $1/a$. A study by Dhrymes and Kurz[34] of the production function of turbogenerators provided some evidence concerning the homogeneity of the production function. Their production function related output (Q), to labor (L), fuel (F), and capital (K) as follows:

$$Q = \min(g(L), (\alpha_f F^{\beta_f} + \alpha_k K^{\beta_k})^{1/\gamma}).$$

In this production function, substitution is possible between fuel and capital but not between those factors and labor. Since labor entered into the production function in such an awkward manner and is a less important factor of production than fuel or capital, accounting for approximately 10 percent of production cost, I ignored it in my further analysis.

Dhrymes and Kurz stratified their sample of 362 generally single turbogenerator plants by time period and size class into thirteen cells. They estimated a relationship between fuel and capital in which the parameter $(\beta_k - 1)/(\beta_f - 1)$ appeared. This parameter equals one only if β_k equals β_f, the homogeneity condition for the production function relating output to fuel and capital. In six cells out of thirteen it was not possible to reject the null hypothesis that the parameter was equal to one using a two-tailed t-test with a 5-percent significance level, although it was true that the parameter was greater than one in eleven cases and less than one in only two cases. In view of the fact that the homogeneity assumption was accepted in half the cells and that the theory was simple with it, I proceeded conditionally on the assumption that it was valid.

In the cells in which $\beta_f = \beta_k$ the economies of scale parameter was simply γ/β_f; in the cells in which $\beta_f \neq \beta_k$ a further calculation was made by Dhrymes and Kurz to derive an approximate economies of scale parameter. The estimates were used to provide prior information for this study as follows.

If the estimates denoted $\hat{a}_i$, $i = 1, 2 \ldots m$, were hypothesized to be drawn from a normal population with mean a and an unknown but constant variance ϕ^2 and if the prior distributions of a and $\log \phi$ were independent and each uniform, then it follows from Bayes' theorem that the posterior joint probability density function is given by the following expression:

$$p(a, \phi | \hat{a}_1 \ldots \hat{a}_m) \propto \phi^{-(m+1)} \exp \left\{ -\frac{1}{2\phi^2} \sum_{i=1}^{m} (\hat{a}_i - a)^2 \right\}.$$

The posterior marginal distribution of a is derived by completing the square on a and integrating with respect to ϕ over the range zero to infinity:

$$p(a | \hat{a}_1, \ldots, \hat{a}_m) \propto \{(m-1)s^2 + m(a - \hat{a})^2\}^{-m/2}$$

$$\hat{a} = \frac{1}{m} \sum_{i=1}^{m} \hat{a}_i, \; s^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{a}_i - \hat{a})^2.$$

There are ten postwar median estimates of $1/a$ reported in Table

---

[34] In [6].

10 of the Dhrymes-Kurz study; these were inverted and $\hat{a}$ and s^2 were evaluated. The resulting probability density function to be used as a prior probability density function in this study is given by the following expression:

$$p(a|\hat{a}_1 \ldots \hat{a}_{10}) \propto \{9 \times 0.559 \times 10^{-2} + 10(a - 0.8974)^2\}^{-5}.$$

This is in the form of a univariate Student t-distribution with 9 degrees of freedom, a mean of 0.8974, and a standard deviation of 0.027.

The prior probability density function used incorporates two separate pieces of information. First, the prior distributions of a, θ, $\log \sigma_\epsilon$, ρ, $M_1 \ldots M_{N_f}$ are assumed to be independent and each uniform except for that of a. Second, it expresses the fact that a zero probability density should be assigned to any pair (a, θ) for which the existence condition of the original cost minimizing problem is not satisfied. The existence condition was that $\theta - ag_\iota > 0$ for all the investments being used to construct the likelihood function. The most important value of the growth rate in the restriction was obviously the largest one, called $\hat{g}_{\max}$.[35] Thus, the prior joint density function is

$$p(a, \theta, \sigma_\epsilon, \rho, M_1 \ldots M_{N_f}) \begin{cases} = 0, \text{ if } \theta - a\hat{g}_{\max} \leqslant 0 \\ \propto \dfrac{1}{\sigma_\epsilon} \times \{9 \times 0.559 \times 10^{-2} \\ \quad + 10(a - 0.8974)^2\}^{-5}, \text{ otherwise.} \end{cases}$$

To obtain the posterior joint density function, the prior joint density function is multiplied by the likelihood function:

$$\begin{aligned} p(a, \theta, \sigma_\epsilon, \rho, M_1 \ldots M_{N_f}|\log \bar{x}_{11}, \log \bar{x}_{12} \ldots) \\ \propto p(a, \theta, \sigma_\epsilon, \rho, M_1 \ldots M_{N_f}) p(\log \bar{x}_{11}, \\ \log \bar{x}_{12} \ldots |a, \theta, \sigma_\epsilon, \rho, M_1 \ldots M_{N_f}). \end{aligned}$$

The nuisance parameters σ_ϵ, ρ, $M_1 \ldots M_{N_f}$ are removed by analytical integration to yield the following joint posterior probability density function of a and θ:

$$p(a, \theta|\text{data}) \begin{cases} = 0, \text{ if } \theta - a\hat{g}_{\max} \leqslant 0 \\ \propto \dfrac{\{9 \times 0.559 \times 10^{-2} + 10(a - 0.8974)^2\}^{-5}}{(\Sigma\Sigma\, y_{\iota n}{}^2{}_{-1})^{1/2} \left(\Sigma\Sigma\, y_{\iota n}{}^2 - \dfrac{(\Sigma\Sigma\, y_{\iota n} y_{\iota n-1})^2}{\Sigma\Sigma\, y_{\iota n}{}^2{}_{-1}}\right)^{\frac{\nu - N_f - 1}{2}}}, \\ \qquad\qquad\qquad \text{otherwise.} \end{cases}$$

The indices on the summation signs are $\displaystyle\sum_{\iota=1}^{N_f} \sum_{n=2}^{n_\iota}$.

The marginal distributions of a and θ were obtained by numerical integration. Preliminary work suggested that it was reasonable to pool observations from different firms. Details of the posterior density functions of a and θ are shown in the first two columns of Table 5.[36] A plot of the posterior joint density function of a and θ

---

[35] The estimated maximum growth rate was used in place of the actual (and unknown) maximum rate of growth.

TABLE 5

SUMMARY STATISTICS OF THE DISTRIBUTIONS OF a AND θ AND RESIDUAL STATISTICS

	SUBSAMPLE 1	SUBSAMPLE 2	SUBSAMPLE 1
DEPENDENT VARIABLE	LOG (ANTICIPATED CYCLE TIME)	LOG (ANTICIPATED CYCLE TIME)	LOG (ACTUAL CYCLE TIME)
ERROR PROCESS	AUTOREGRESSIVE	AUTOREGRESSIVE	AUTOREGRESSIVE
NUMBER OF OBSERVATIONS*	48	62	42
JOINT P D F OF a, θ			
MODE	(0 88, 0 10)	(0 90, 0 09)	(0 88, 0 10)
MARGINAL P D F OF a			
MEAN	0 8716	0 8772	0 8668
STANDARD DEVIATION	0 0274	0 0222	0 0200
MARGINAL P D F OF θ			
MEAN	0 1455	0 1239	0 1215
STANDARD DEVIATION	0 0564	0 0368	0 0254
MODE	0 100	0 091	0 108
ESTIMATED SERIAL CORRELATION COEFFICIENT**	0 90	0 71	0 24

* THERE ARE FEWER OBSERVATIONS WHEN THE DEPENDENT VARIABLE IS ACTUAL CYCLE TIME, BECAUSE AN ANTICIPATED CYCLE TIME CAN BE CONSTRUCTED FOR THE LAST UNSHARED INVESTMENT OF EACH FIRM BEFORE 1968

** THE COMPUTATION OF THIS WAS BASED ON THE MEAN RESIDUALS

for subsample two is presented in Figure 1. Finally a plot of the mean residuals for subsample two is presented in Figure 2.[37]

The similarity of the results for subsamples one and two is striking. What is disappointing is the considerable serial correlation of the mean residuals, the large standard deviations and the somewhat large means of the posterior distributions of θ especially for subsample one. Having had some time to reflect on these results, I now believe that my decision to use the anticipated cycle time series rather than the actual cycle time series as the dependent variable in equation (12) induced spuriously much of the serial correlation of the errors; this in turn caused the posterior distributions of θ to have large standard errors and this caused the means of these distributions to be biased upward.

Recall that the anticipated cycle time series was constructed by using the estimated rate of growth of peak load. Suppose that

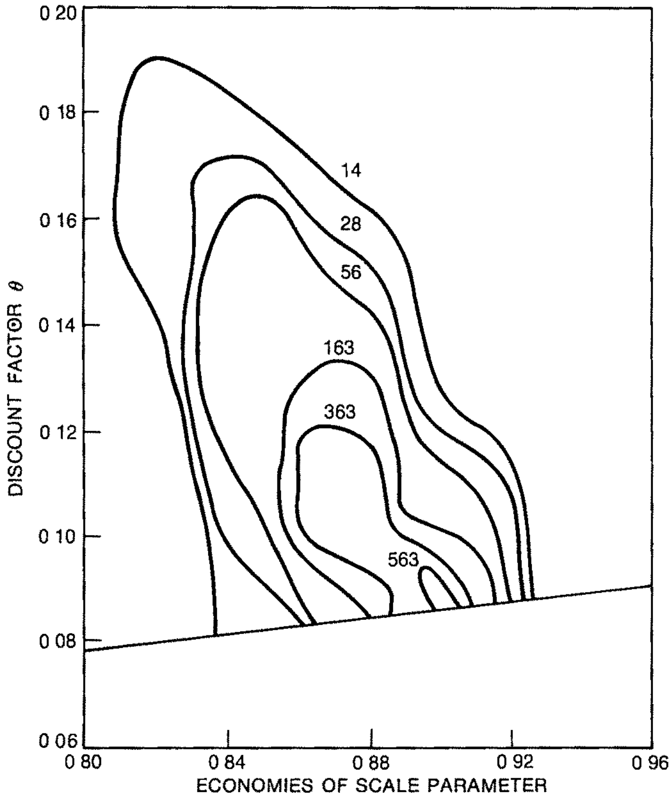
---

[36] The investments carried out by Dallas and Tampa during the period 1947 through 1958 were excluded from the computation of the probability density function for subsample two. The two firms were growing at such a rapid rate in this period that it seemed that the existence condition for a solution, $\theta - ag > 0$, could not be met and thus that the investments of these firms could not conform to the modified d'Aspremont model. During the period 1947 through 1958 the peak load of Dallas was growing at an estimated rate of 0 1425 per annum and that of Tampa at 0.1445.

[37] The mean residuals $\overline{w}_{tn}$ were constructed similarly to the mean residuals $\overline{v}_{tn}$. See note 30.

FIGURE 1

CONTOURS OF THE JOINT DENSITY FUNCTION OF a AND θ FOR FIRMS IN SUBSAMPLE TWO



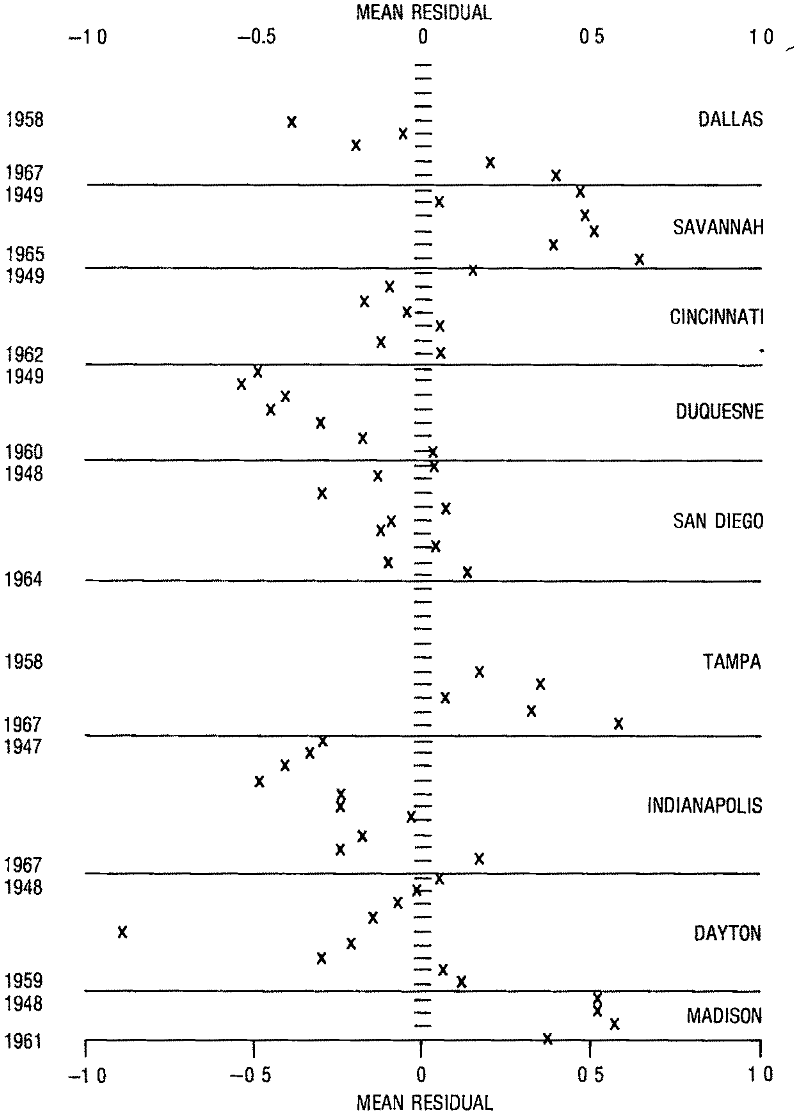
NOTE FOR SUBSAMPLE TWO, THE LARGEST ESTIMATED GROWTH RATE WAS 0 0973 THE LINE CONNECTING THE TWO (a, θ) POINTS (0 8, 0 078) AND (0 96, 0 093) IS THE LINE θ −0 0973a = 0 BELOW THIS LINE, THE POSTERIOR JOINT DENSITY IS ZERO IT IS APPARENT THAT THIS CAUSES CONSIDERABLE TRUNCATION OF THE DISTRIBUTION

for a ten-year period, the manager of a firm anticipated that peak load would grow at 10 percent per annum; suppose too that I had estimated the annual rate of growth to be 8 percent. Then every anticipated cycle time I constructed for this firm during the ten-year period would be too long. This could easily cause much of the serial correlation observed. Next it is well known, at least in the case of linear regression, that a considerable degree of serial correlation will cause the posterior distributions of the parameters to be somewhat diffuse, hence the high standard deviations of the distributions of θ. Finally, the posterior distributions of θ are truncated for low values of θ due to the existence conditions that $\theta > ag$; this is apparent from Figure 1. A large standard deviation coupled with truncation of the lower tail of the distribution will lead to a high mean. As evidence in favor of this interpretation, the posterior distribution of a and θ was computed for the firms in subsample one by using actual rather than anticipated cycle time as the dependent variable. The results are displayed in column 3 of Table 5. The mean and standard deviation of θ are considerably lower.

The means of the posterior distributions of the discount rate θ were about 12 percent and their modes were around 10 percent.

FIGURE 2

PLOT OF MEAN RESIDUALS FOR INVESTMENTS OF FIRMS IN SUBSAMPLE TWO



NOTE THIS FIGURE DISPLAYS THE MEAN RESIDUALS FOR EACH INVESTMENT, THEY ARE PLOTTED SO THAT FOR EACH COMPANY, THE LOWER A POINT IS, THE LATER IS THE TIME OF THE INVESTMENT TO WHICH IT REFERS.

This is higher than one would expect. Recall that the discount rate θ should be interpreted as the sum of the real cost of capital r and the anticipated rate of technical change ψ. The work of Miller and Modigliani[38] provides estimates of the cost of capital for three years in the 1950s. They are 0.036 for 1954, 0.045 for 1956, and 0.046 for 1957. If the assumption is made that no inflation was anticipated in the 1950s, these estimates can be interpreted as estimates of the real cost of capital. Thus, r is approximately 4 percent. The work of Belinfante[39] provides estimates of the actual rate of embodied technical change. He found that the estimated annual rate of technical change of the production function for the period 1947 through 1958 was 0.0309 with a standard error of 0.0138. If the economies of scale parameter were exactly 0.9, a rate of technical change of the production function of 0.0309 would imply a rate of technical change of the cost function of 0.0278. Thus ψ is approximately 3 percent. So, we should expect the discount factor θ to be around 7 percent. In further work, this discrepancy between 7 percent and 12 percent (the mean) or 10 percent (the mode) should be evaluated as follows.

---

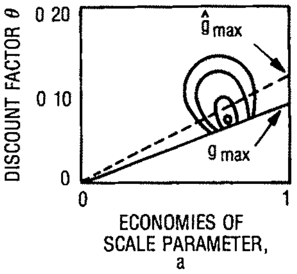
[38] In [21].

[39] In [2].

An investigation should be made concerning the effect of using the estimated maximum rate of growth as a proxy for the actual maximum rate of growth in the existence condition $\theta - ag_{\max} > 0$. It will be recalled that this condition imposes a strong restriction on the parameter space by specifying that the posterior density be zero for any combination (a, θ) for which $\theta - ag_{\max} \leqslant 0$. Figure 3 shows the contours of the posterior probability density function of (a, θ) for the true maximum growth rate $g_{\max}$. All points below the line from 0 to $g_{\max}$ have a zero probability density. Now suppose that $\hat{g}_{\max}$ is used to estimate $g_{\max}$. Since $\hat{g}_{\max}$ is the highest estimated growth rate, it is most likely to be an overestimate. Then all points below the line 0 to $\hat{g}_{\max}$ have a zero probability density and the mean and possibly the mode of the posterior distribution of θ will be biased up.

FIGURE 3

THE EFFECT OF OVERESTIMATING $g_{\max}$



10. Measures of fit for the modified d'Aspremont model

■ Exact posterior probability distributions of the parameters have been provided together with residual summary statistics and plots. But this does not provide a real sense of how well the model actually fits the data. This section develops two measures of the fit of actual aggregate investment to investment predicted by the model during the period 1950 through 1966. These measures are later compared with a measure of the fit of a distributed lag model estimated for aggregate investment over the same period 1950 through 1966. The period of comparison began in 1950 to provide sufficient initial values and ended in 1966 because this was the last year before any company in the sample engaged in joint installation.

The first type of prediction is designed to test how well the model predicts the magnitude of aggregate investment; this will be referred to as a conditional prediction. The second type of prediction is designed to test how well the model predicts both the magnitude of aggregate investment and also its timing; this will be referred to as a one-year-ahead prediction.

An example is now provided of the derivation of the predictions; for the sake of clarity, Madison will be discussed. First, for each of the years 1950 through 1966, the turbogenerator size implied by the model, if an investment had been made, was computed, according to equation (13):

$$z_t^* = 1 + \bar{\delta}/\hat{g})\, \hat{D}_t\, (e^{\hat{g}x_t^*} - 1) \qquad t = 1950 \ldots 1966, \quad (13)$$

where

$$x_t^* = \frac{2(1 - \bar{a})}{\hat{g}(1 - \bar{a}^2) + \bar{a}\,\bar{\theta}}$$

and

$$D_t = \exp(\hat{\alpha} + \hat{g}(t - 1946.5) + 1/2\, \hat{\sigma}_u^2).$$

The parameters a, θ, and δ are surmounted by a bar to signify that the means of the posterior distributions are being used in these expressions. Also the company index i has been dropped to simplify the notation.

To make conditional predictions for Madison, one notes that Madison made investments in 1953, 1957, and 1961. Madison's conditional prediction in each of these three years is set to whatever the value of z^* is in each year; in all other years in the period 1950 through 1966 the conditional prediction is set to zero. Conditional predictions were made this way for all the other firms and the aggregate conditional prediction in any year is simply the sum of all fifteen firms' conditional predictions.

One-year-ahead predictions are made conceptually on December 31st of the preceding year. Again, consider Madison which made its second investment of size $z_2 = 30$ MW. on September 23, 1953, and made its third investment in 1957. The actual time between the second investment and the third is a random variable, the distribution of which is given by equation (14):

$$X_2 = \frac{1}{\hat{g}} \log\left(1 + \frac{z_2}{\hat{D}_{1954}\,(1 + \bar{\delta}/\hat{g})}\right) e^{v_2}, \qquad (14)$$

where

$$v_2 \sim N(0, \sigma_v^2).$$

From this equation, it is possible to compute the probability that there is another investment in the remainder of 1953, that there is one in 1954, that there is one in 1955 and so on.

As of December 31, 1953, it is known that no further investment took place in 1953 and the probability of observing an investment in 1954 is now recomputed according to the formula

$$\text{probability}\begin{pmatrix}\text{investment in 1954}\\ \text{given none in the}\\ \text{remainder of 1953}\end{pmatrix} = \frac{\text{probability (investment in 1954)}}{1 - \text{probability (investment in 1953)}}. \qquad (15)$$

The one-year-ahead prediction of investment for 1954 is derived by multiplying the probability derived in expression (15) by z^* for 1954 computed according to equation (13).

Now we move on one year. As of December 31, 1954, it is known that there was no investment in 1954 and by using an expression analogous to (15) a new probability is computed that an investment would take place in 1955, given that there had been none in the remainder of 1953 nor in 1954. This probability is multiplied by z^* for 1955, computed according to equation (13). This procedure of recomputing the probabilities and multiplying them by z^* for the appropriate year continues until 1957 when an investment really did take place, at which time we start all over again with an expression analogous to (14). One-year-ahead predictions were made this way for all fifteen firms and the aggregate one-year-ahead prediction in any year is the sum of each firm's prediction in that year.[40]

Summary statistics of the fit of predicted to actual aggregate investment were produced as follows. Let $\widehat{\Delta C}_{\cdot t}$ represent actual aggregate investment in year t and $\Delta C_{\cdot t}$ represent its prediction. Let T_i represent the initial year of the comparison and T_f represent the final year. Then the standard error of predicted investment is defined as

$$s_p = \left(\frac{1}{T_f - T_i} \sum_{t=T_i}^{T_f} (\widehat{\Delta C}_{\cdot t} - \Delta C_{\cdot t})^2 \right)^{1/2} .$$

Standard errors of prediction for various periods are displayed in Table 6. In addition, plots are provided in Figure 4 of actual aggregate investment, the conditional predictions and the one-year-ahead predictions. As expected the conditional predictions are more precise than the one-year-ahead predictions.

TABLE 6

STANDARD ERRORS OF PREDICTION OF THE d'ASPREMONT MODEL AND THE DISTRIBUTED LAG MODEL

	1950-1960	1961-1966	1950-1966
CONDITIONAL PREDICTIONS FROM d ASPREMONT MODEL	109	170	128
ONE-YEAR-AHEAD PREDICTIONS FROM d'ASPREMONT MODEL	184	517	324
DISTRIBUTED LAG MODEL ($\delta = 0.00$)	290	130	240

It is apparent that the fit of the model is much worse in the latter part of the period, particularly for the one-year-ahead predictions. A glance at Table 2 will show that in the period 1961 through 1966 there were a number of abnormally long intervals between investments, and therefore, few investments made in that period. Consequently the long cycle times in the 1960s did not affect the posterior distributions of the parameters considerably.

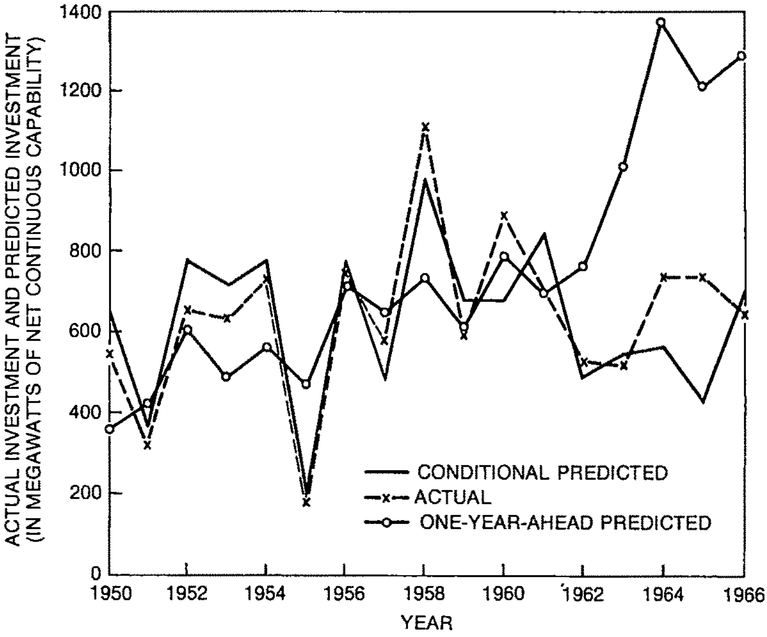
As was pointed out previously, the probability distribution of the actual cycle time is given by an expression analogous to (14). We now show the effect of abnormally long cycle times on one-year-ahead predictions. Consider Indianapolis, which made an investment on May 15, 1961, and then none until 1967. As of May 16, 1961, the probability that an investment would take place in 1965, for example, was 0.0157. But by December 31, 1964, the prob-

---

[40] The numerical values used to make the predictions were as follows. For investments of firms in subsample one, $\bar{a} = 0.8668$, $\bar{\theta} = 0.1215$, $\bar{\delta} = 0.0217$, $\hat{\sigma}_v^2 = 0.27$; for subsample two, $\bar{a} = 0.8772$, $\bar{\theta} = 0.1239$, $\bar{\delta} = 0.0246$, $\hat{\sigma}_v^2 = 0.14$ for all firms except Dayton and $\hat{\sigma}_v^2 = 1.27$ for Dayton. For the source of these values see Tables 4 and 5 and note 2 of Table 4. For Dallas and Tampa, the predictions were made by using the parameters of their peak load growth paths estimated for the period 1958 through 1968. For Cincinnati, the predictions were made by using the parameters of its peak load growth estimated for the period 1953 through 1968.

FIGURE 4

PREDICTIONS MADE WITH THE MODIFIED d'ASPREMONT MODEL



ability that there would be an investment in 1965 had been transformed to 0.83 by repeated application of operations as in expression (15). Thus z^* for 1965 was multiplied by a probability which was close to one. Likewise, as of May 16, 1961, the probability that an investment would take place in 1966 was 0.0026. But by December 31, 1965, the probability had been transformed to 0.82. So z^* for 1966 was multiplied by a probability that was close to one. This explains why one-year-ahead predictions were so high in the 1960s. The more fundamental question as to why cycle times were so long in the 1960s is more tricky. There were six firms which did not invest much in the 1960s—Madison, Illinois, Indianapolis, Dayton, Duquesne, and Cincinnati. Except for Illinois, all these companies made abnormally large imports of electricity in the 1960s, so they were clearly not self-sufficient. In addition, four of them—Madison, Dayton, Duquesne, and Cincinnati—were engaged in joint planning with other firms toward the end of the 1960s, so the modified d'Aspremont model could not be expected to apply to these firms alone, but to the entire pool to which they belonged.

Moreover, the use of gas-turbine peaking units became common in the 1960s. It is my impression that in the 1950s most firms had been purchasing conventional steam-electric turbogenerator units both for base-load and peak-load purposes; they had, therefore, operated their steam-electric capacity at a relatively low average rate of utilization. In the 1960s gas-turbine units became generally available and many firms installed them for peak load purposes. They then found that they had too much steam-electric capacity since they only had to use this for base-load purposes. They therefore waited until the rise in demand had increased the rate of utilization of their steam-electric capacity. This caused cycle times to increase in the 1960s.[41]

11. The fixed lag model

The fixed lag model of investment for a firm was shown to be expressed by the following equation:

$$\Delta C_{it} = \Delta D_{it-\tau} + \delta C_{it} + u_{it}.$$

The subscript i has been added to refer to firm i. The only unknown parameter is τ, the order of the lag.

It was pointed out previously that an implication of this model is that if demand increased almost monotonically, then there should be an investment in almost every year; this implication is strengthened by the reasonable condition that the rate of depreciation be positive. It is apparent from Table 7 that in fact investments occur approximately only half as frequently as increments in demand, whether demand is measured by peak load or by annual net generation.[42] This evidence refutes the fixed lag model of investment for this body of data.

TABLE 7

COMPARATIVE FREQUENCY OF INCREMENTS OF DEMAND AND INCREMENTS OF CAPACITY

FIRM	PROPORTION OF YEARS IN WHICH PEAK LOAD RISES 1945-1966	PROPORTION OF YEARS IN WHICH NET GENERATION RISES 1945-1966	PROPORTION OF YEARS IN WHICH THERE ARE INVESTMENTS 1947-1968
MADISON	15/16[a]	17/18	4/21
SAVANNAH	14/17[a]	21/22	6/22
EL PASO	21/22	22/22	8/22
ILLINOIS	19/22	20/22	5/22
NEW ORLEANS	16/22	16/22	8/22
INDIANAPOLIS	20/22	17/22	9/22
DAYTON	22/22	19/22	8/22
TAMPA	22/22	21/22	10/22
LOUISVILLE	20/22	14/22	9/22
SAN DIEGO	21/22	22/22	10/22
BOSTON	18/22	19/22	8/22
DUQUESNE	20/22	17/22	9/22
CINCINNATI	11/13[a]	19/22	9/22
BALTIMORE	18/22	16/22	12/22
DALLAS	22/22	22/22	13/22

[a] MADISON'S PEAK LOAD WAS REPORTED FROM 1950 ON, WHEREAS ITS NET GENERATION AND INVESTMENTS WERE REPORTED FROM 1948 ON SAVANNAH'S PEAK LOAD WAS NOT REPORTED IN THE YEARS 1959-1962 INCLUSIVE CINCINNATI'S PEAK LOAD WAS NOT REPORTED UNTIL 1953

---

[41] The technology of gas-turbine peaking units is different from that of conventional steam-electric units. The gas-turbine units have a relatively high ratio of variable to fixed costs and are relatively small. It was not possible to incorporate the problem of the optimal mix of plants as between conventional and peaking into the modified d'Aspremont model and retain the model's tractability. Therefore, the model was interpreted as explaining only investments in conventional capacity and the introduction of peaking units is being used here to suggest why the model went astray in the 1960s.

[42] The table was constructed on the assumption of a two-year lag. Thus the number of investments over the period 1947 through 1968 is being compared with the number of increments of demand over the period 1945 through 1966. I do not feel that changing the order of the lag would affect the conclusions materially. Peaking units were not counted as an investment when the table was constructed.

12. The distributed lag model

■ It has been shown that the distributed lag model of firm investment reduces to a fixed lag model when explaining lumpy investments, and furthermore that the fixed lag model is refuted for this body of data.

Nevertheless a distributed lag model of investment in turbogenerator sets by all fifteen firms was estimated because it provided a benchmark for judging the fit of the modified d'Aspremont model, and because such models of aggregate investment are used so widely.

Time series were constructed of annual aggregate investment, annual peak load and annual capacity for each of the years 1947 through 1966.[43] If aggregate investment is denoted by $\Delta C_{\cdot t}$, aggregate peak load by $\Delta D_{\cdot t}$, and aggregate capacity by $C_{\cdot t}$, the distributed lag model is[44]

$$\Delta C_{\cdot t} = \mu(\theta)\, \Delta D_{\cdot t} + \delta C_{\cdot t} + u_{\cdot t},$$

where

$\mu(\theta)$ is a polynomial in the backshift operator θ

$$\mu(\theta) = \mu_0 + \mu_1\theta + \mu_2\theta^2 + \dots.$$

Since there were only 20 observations of aggregate investment running from 1947 through 1966, no attempt was made to estimate the rate of depreciation, δ, as well as the parameters of the lag distribution. Instead the model was estimated twice, once assuming that $\delta = 0.00$ and once assuming that $\delta = 0.03$.

To represent the lag distribution, $\mu(\theta)$, Jorgenson's rational distributed lag technique[45] was employed. The following specifications of the lag distribution were tested:

$$\frac{1-b}{1-b\theta}, \frac{1-b}{1-b\theta}\theta, \frac{1-b-c}{1-b\theta-c\theta^2}, \frac{1-b-c}{1-b\theta-c\theta^2}\theta.$$

In choosing the best specification of the lag function, the following criteria were used:

(1) If the Durbin-Watson statistic was greatly different from 2, that lag specification was rejected. This was especially important in view of the fact that the D-W statistic is biased toward a value of 2 in an autoregressive model.

(2) Lag specifications were rejected in which some of the parameters were not significantly different from zero at the 5-percent significance level. With such a small number of observations, this induced a predisposition to accept models with few parameters.

(3) Specifications with a constant term were rejected, if the constant term was not significantly different from zero at

---

[43] Jorgenson, in his estimation of distributed lag models prefers to use quarterly investment data. This was not feasible here, since with only fifteen companies, the quarterly series would not have been sufficiently smooth.

[44] The aggregate investment series was composed only of those investments with which the modified d'Aspremont model had been estimated; thus, no peaking units were included. Investment, capacity, and peak load were all measured in megawatts.

[45] In [12].

the 5-percent level. Since the model does not imply a constant term, it did not seem satisfactory to include such a term unless it was really necessary.

Given these criteria, the best results of the distributed lag estimation were the following first-order schemes, which were estimated by ordinary least squares. The quantities in parentheses are the standard deviations of the least squares estimators.

For $\delta = 0.00$,

$$(\Delta C_{\cdot t} - \Delta D_{\cdot t-1}) = \underset{(0.1618)}{0.4621}\,(\Delta C_{\cdot t-1} - \Delta D_{\cdot t-1}) + \hat{u}_{\cdot t}$$

$$s = 243.9 \qquad D.W. = 2.03,$$

implying a lag distribution $\mu(\theta) = \dfrac{0.5379}{1 - 0.4621\theta}\,\theta.$

For $\delta = 0.03$,

$$(\Delta C_{\cdot t} - .03C_{\cdot t} - \Delta D_{\cdot t-1}) = \underset{(0.1523)}{0.5798}\,(\Delta C_{\cdot t-1} - .03C_{\cdot t-1} - \Delta D_{\cdot t-1}) + \hat{u}_{\cdot t}$$

$$s = 254.9 \qquad D.W. = 2.21,$$

implying a lag distribution $\mu(\theta) = \dfrac{0.4202}{1 - 0.5798\theta}\,\theta.$

It is well known that the results of the least squares estimation can be given a Bayesian interpretation. If the prior distributions of b and $\log \sigma_u$, where $u_{\cdot t}$ is the equation error, are independent and each uniform, then the posterior distribution of the parameter b is univariate Student t with a mean numerically equal to the least squares estimate and a standard deviation numerically equal to the sampling theory standard error for the regression coefficient.[46]

For each of the years 1950 through 1966, for the zero rate of depreciation model, fitted investment was constructed conditional on the mean of the posterior distribution of b. Thus, fitted investment was constructed by means of the relationship

$$\widehat{\Delta C}_{\cdot t} = 0.4621\ \Delta C_{\cdot t-1} + 0.5379\Delta D_{\cdot t-1}. \qquad (16)$$

Fitted investment in year t can be thought of as a prediction made on December 31st of the preceding year. Both fitted investment and actual investment are plotted in Figure 5. It is apparent that fitted investment does not conform very closely to actual investment. Actual investment is considerably more volatile than fitted investment. The standard error of prediction was computed for the distributed lag model just as it had been for the modified d'Aspremont model and it is presented in Table 6.

13. A comparison of the modified d'Aspremont model and the distributed lag model

■ The modified d'Aspremont model was found to be consistent with the individual firm data whereas the fixed lag model was not. The comparison of the models using the aggregate data was not so straight-forward.

The most relevant comparison is between the one-year-ahead predictions made with the modified d'Aspremont model and the fitted values of investment from the distributed lag model. This is because they may both be viewed as predictions made on the same date—December 31 of the preceding year. From Table 6 it is apparent that the d'Aspremont model dominates the distributed lag model for the period 1950 through 1960, but is dominated by it for the period 1961 through 1966. The reason for the bad prediction performance of the d'Aspremont model in the 1960s has already been discussed.

---

[46] See Zellner [33], pp. 66–67.

A less relevant comparison, although still interesting is between the conditional predictions made with the modified d'Aspremont model and the fitted values of investment of the distributed lag model. The conditional predictions are made "later" than the distributed lag predictions in the sense that they are made conditional on the knowledge that an investment actually took place in a particular year. The former predictions should therefore be better. This is indeed the case for the period 1950 through 1960, as is apparent from Table 6. It is slightly disturbing that the distributed lag predictions are better for the period 1961 through 1966.

14. Conclusions and implications for further work

■ It has been demonstrated that it is sometimes useful to analyze models by using disaggregated data. For instance, it was quite clear that the fixed lag model was not consistent with the individual firm data, so that it was possible to make a clear choice between the two competing models of firm investment. In addition, predictions of aggregate investment made with the modified d'Aspremont model were more successful than those made with the distributed lag model for the period 1950 through 1960, although the ranking was reversed in the period 1961 through 1966. But the latter finding is not too dismaying. After all, something unusual was happening in the period 1961 through 1966, and this was not apparent from the fit of the distributed lag model.

Further studies are either planned or under way. Their variety should suggest the fruitfulness of lumpy models of investment. The first concerns the modeling of an industry in which each firm makes lumpy investments. Preliminary work suggests that it is not clear *a priori* whether the aggregate investment series will be sufficiently smooth to be fit by a distributed lag model as was suggested by Koyck.[47]

Consider an "industry" composed of two firms. Each firm's demand grows arithmetically at an annual rate of 50 megawatts and there is no depreciation. Then the distributed lag model would predict that industry investment in the steady state should be 100 megawatts per year. Suppose that each firm's optimal policy is to make an investment of 100 megawatts every two years. Then if the two firms invest synchronously, industry investment will be 200 megawatts one year, none the next, 200 megawatts the next, and so on. If the firms invest asynchronously, industry investment will be 100 megawatts a year as predicted by the distributed lag model.

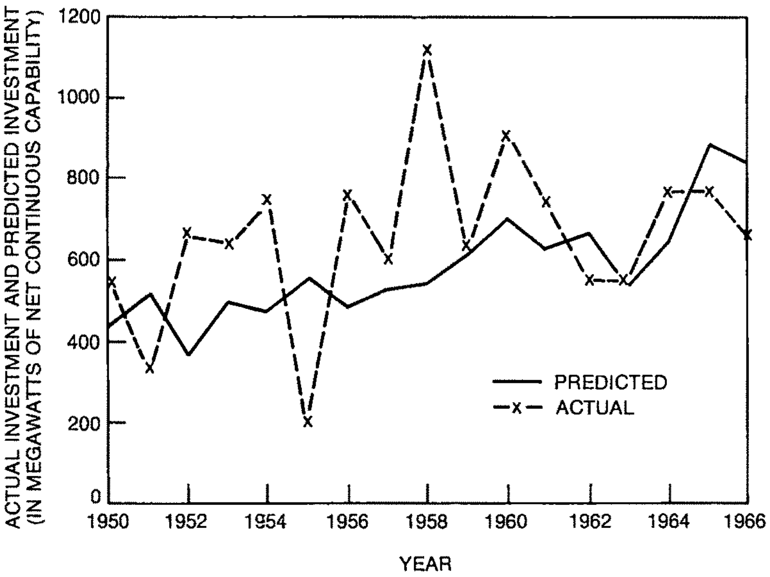
It appears that an important parameter in this exercise is the ratio of the number of firms in the industry to the cycle time. As the number of firms rises, one can presumably make the argument that firms are more likely to invest asynchronously and thus that the aggregate investment series should become smooth. This argument may not be correct; from Figure 5 it is apparent that the investment of an "industry" of fifteen firms with an average cycle time of two years was not particularly smooth. The aggregate investment series may be smoothed if investment is measured as dollar outlays rather than megawatts of installed capacity, since it is common in the industry to make payments while the turbogenerator set is being fabricated and installed. This is clearly an important question for investigation. Simulation methods will be used to construct the aggregate investment of an industry composed of firms making lumpy investments and the smoothness of the aggregate series will be examined.

---

[47] In [16].

FIGURE 5

PREDICTIONS MADE WITH THE DISTRIBUTED LAG MODEL ($\delta = 0\ 00$)



Simulation methods will also be used to examine the response of an industry to exogenous changes such as the imposition of an excise tax. The tax will presumably increase output price, which will cause a slowing down of demand growth for each firm and hence a postponement of the next investment. Thus an excise tax should temporarily reduce the rate of investment in the industry; this in itself is not a particularly surprising result.[48] What is interesting is the time pattern of the response which will be yielded by the simulations.

A second study concerns the effect of the formation of power pools in the electric utilities industry on new investment. In the 1960s many firms formed power pools and began to plan the installation of jointly owned turbogenerators. The presumption is that the new turbogenerator sizes were planned on the basis of the demand growth of the power pools rather than of the individual firms. Data have been gathered on the joint installations already made and planned by a number of separate power pools. Moody's *Public Utilities, 1972*, reports installation plans until 1977 or 1978. An investigation will be made as to whether the investments of the pools conform to the modified d'Aspremont model or a variant of it. A cursory inspection of the data reveals that the joint installations of a power pool are typically much larger than would be made by any one member of the pool, which is what the model would lead one to expect. But there is an additional problem with this body of data which was not present in the preceding study. It will be recalled that the study was concerned with the investments of moderately sized to small firms. So that, for instance, if the optimal cycle time was two years, there was a presumption that a turbogenerator existed which was large enough to meet the increased demand and the depreciation for a two-year period. Power pools are typically large, however, and turbogenerators large enough for a two-year period may not exist. Therefore, there may be an identification problem in that it may not be possible to make inferences about a and θ from the investment behavior of the power pools.

---

[48] Some preliminary work suggests that the steady state rate of investment may be reduced by the imposition of the excise tax.

A final study which is planned may resolve a paradox in the field of comparative industrial organization. Bain[49] found that in Japan, France, Italy, and India multiplant operation by leading firms was about as prevalent as it was in the United States; it was somewhat lower in England and substantially lower in Canada and Sweden. However, except for England, plant sizes abroad measured in terms of employment were much smaller on the average than in the United States. Now this finding is certainly consistent with a model of the d'Aspremont type. If a firm has a small market or grows slowly or both, then we should expect it to build small plants, given that the economies of scale are embodied in the plants. Furthermore, the total market in a country like Italy is small in comparison with the market in the United States. Therefore, it is reasonable to assume that firm size and, consequently, plant size will be smaller in Italy than in the United States.

References

1. Bain, J. S. *International Differences in Industrial Structure: Eight Nations in the 1950s.* New Haven: Yale Univ. Press, 1966.
2. Belinfante, A. E. E. "Technical Change in the Steam Electric Power Generating Industry." Working Paper No. 156 in Economic Theory and Econometrics, University of California, Berkeley, September 1969.
3. Capron, W. M., ed. *Technological Change in Regulated Industries.* Washington. D. C.: The Brookings Institution, 1970.
4. Chenery, H. B. "Overcapacity and the Acceleration Principle." *Econometrica,* Vol. 20 (January 1952), pp. 1–28.
5. d'Aspremont, C. Gabszewicz, J. J., and Vial, J. P. "Capital Depreciation, Growing Demand, and Economies of Scale." Core Discussion Paper No. 6918, Université Catholique de Louvain, June 1969.
6. Dhrymes, P. J. and Kurz, M. "Technology and Scale in Electricity Generation." *Econometrica,* Vol. 32, No. 3 (July 1964), pp. 287–314.

---

[49] In [1].

7. Eisner, R. "A Distributed Lag Investment Function." *Econometrica*, Vol. 28 (1960), pp. 1–29.
8. Gould, J. P. "The Use of Endogenous Variables in Dynamic Models of Investment." *Quarterly Journal of Economics*, Vol. 83 (1969), pp. 580–599.
9. Griliches, Z. "Distributed Lags: A Survey," *Econometrica*, Vol. 35, No. 1 (January 1967), pp. 16–49.
10. Grunfeld, Y and Griliches, Z. "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics*, Vol. 42 (February 1960), pp. 1–13.
11. Jorgenson, D. W. "Anticipations and Investment Behavior," in J. S. Duesenberry, G. Fromm, L. R. Klein, and E. Kuh, eds., *The Brookings Quarterly Model of the United States*, Chicago: Rand McNally, 1965, pp. 35–94.
12. ———. "Rational Distributed Lag Functions." *Econometrica*, Vol. 32 (January 1966), pp. 135–149.
13. ——— and Handel, S. S. "Investment Behavior in U.S. Regulated Industries." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 213–264.
14. Kirchmayer, L. K., Mellor, A. G., O'Mara, J F., and Stevenson, J. R. "An Investigation of the Economic Size of Steam-Electric Generating Units." *AIEE Transactions Part III* (August 1955), pp 600–614.
15. Kisselgoff, A. and Modigliani, F. "Private Investment in the Electric Power Industry and the Acceleration Principle," in R. A. Gordon and L. R. Klein, eds., *AEA Readings in Business Cycles*, Vol. 10, Homewood, Ill.: Richard D. Irwin, 1965, pp. 340–368.
16. Koyck, L. M *Distributed Lags and Investment Analysis.* Amsterdam: North Holland, 1963.
17. Lindley, D. V. *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2. Inference.* Cambridge: Cambridge Univ. Press, 1965.
18. Manne, A. S "Capacity Expansion and Probabilistic Growth." *Econometrica*, Vol. 29, No. 4 (October 1961), pp. 632–649.
19. ———. *Investments for Capacity Expansion.* Cambridge: MIT Press, 1967.
20. Massé, P. and Gibrat, R. "Application of Linear Programming to Investments in the Electric Power Industry." *Management Science* (January 1957), pp. 149–166.
21. Miller, M. H. and Modigliani, F. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57." *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 333–391.
22. *Moody's Public Utility Manual.* New York: Moody's Investors Services, Inc., 1946–1969.
23. Peck, S. C. "Tests of Alternative Models of Investment for the Electric Utilities Industry." Unpublished Ph.D. dissertation, University of Chicago, 1973.
24. Pratten, C. F., Dean, R. M., and Silbertson, A. *The Economics of Large Scale Production in British Industries; an Introductory Study.* Occasional Papers No. 3, Department of Applied Economics, University of Cambridge, Cambridge Univ. Press, 1965.
25. Scherer, F. M. "The Determinants of Industrial Plant Sizes in Six Nations." *Review of Economics and Statistics* (May 1973), pp. 135–145
26. Scitovsky, T. *Economic Theory and Western European Integration.* Stanford: Stanford Univ. Press, 1958.
27. Srinivasan, T. N. "Geometric Rate of Growth in Demand," in A. S. Manne, ed., *Investments for Capacity Expansion*, Cambridge: MIT Press, 1967.
28. U. S. Federal Power Commission. *Electric Power Statistics.* Washington, D. C.: U. S. Government Printing Office, 1969.
29. ———. *Glossary of Important Power and Rate Terms, Abbreviations, and Units of Measurement.* Washington, D C.: U. S. Government Printing Office, 1965.

30. ———. *Statistics of Electric Utilities in United States, Classes A and B Privately Owned Companies.* Washington, D. C.: U. S. Government Printing Office, 1938–1969.
31. ———. *Steam Electric Plants: Construction Cost and Annual Production Expenses.* Washington, D. C.: U. S. Government Printing Office, 1938–1969.
32. VENNARD, E. *The Electric Power Business.* New York: McGraw-Hill, 1970.
33. ZELLNER, A. *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley & Sons, Inc., 1971.

SYMPOSIUM ON ANALYTICAL MODELS OF ENERGY POLICY

461 U. S. energy policy and economic growth, 1975-2000
E. A. HUDSON and D. W. JORGENSON

515 The effects of higher prices on electricity consumption
J. M. GRIFFIN

540 An economic model of the world oil market
M. KENNEDY

578 Energy, environment, and economic growth
A. P. CARTER

U. S. energy policy and economic growth, 1975-2000

Edward A. Hudson

Senior Economist
Data Resources, Incorporated

and

Dale W. Jorgenson

Professor of Economics
Harvard University

This paper presents a new approach to the quantitative analysis of U.S. energy policy, based on an integration of econometric modeling and input-output analysis. It incorporates a new methodology for assessing the impact of economic policy on both demand and supply for energy within a complete econometric model of the U.S. economy. The model consists of production models for nine industrial sectors, a model of consumer demand, and a macro-econometric growth model for the U.S. economy. The model is first used to project economic activity and energy utilization for the period 1975 to 2000 under the assumption of no change in energy policy. The model is then employed to design a tax program for stimulating energy conservation and reducing dependence on imported sources of energy. The overall conclusion of the analysis of tax policy is that substantial reductions in energy use can be achieved without major economic cost.

1. Introduction

■ The dramatic increase in world petroleum prices associated with the Arab oil embargo of October 1973 has highlighted the need for a new approach to the quantitative analysis of economic policy. Econometric models in the Tinbergen-Klein mold have proved to be very useful in studying the impact of economic policy on aggregate demand.[1] At the same time these models do not provide an adequate basis for assessing the impact of economic policy on supply. Input-output analysis in the form originated by Leontief is useful for a very detailed analysis of supply, predicated on a fixed technology at any point of time.[2] Input-output analysis does not provide

---

Dale W. Jorgenson received the B.A. from Reed College (1955), and the A.M. and Ph.D. from Harvard University (1957 and 1959, respectively). He was the 1971 recipient of the John Bates Clark Medal of the American Economic Association. His interests include the econometrics of production and investment.

Edward A. Hudson received the B.A. from Victoria University, New Zealand, in 1968 and the Ph.D. from Harvard University in 1973. His research interests include energy economics and public finance.

[1] The seminal contribution to macro-econometric modeling of the U. S. economy is the Klein-Goldberger model in [27]. For a recent review of macro-econometric models of the United States, see Hickman [23].

[2] For the original development of input-output analysis, see Leontief [30].

a means of assessing the impact of changes in technology induced by price variations associated with changes in economic policy.

The purpose of this paper is to present a new approach to the quantitative analysis of U. S. energy policy.[3] This approach is based on an integration of econometric modeling and input-output analysis and incorporates an entirely new methodology for assessing the impact of economic policy on supply. We combine the determinants of energy demand and supply within the same framework and relate patterns of U. S. economic growth to both demand and supply. Our approach can be used to project U. S. economic growth and energy utilization for any proposed U. S. energy policy. It can be employed to study the impact of specific policy changes on energy demand and supply, energy price and cost, energy imports and exports, and on U. S. economic growth.

The first component of our framework for energy policy analysis is an econometric model of interindustry transactions for nine domestic industries. We have subdivided the business sector of the U. S. economy into nine industrial groups in order to provide for the detailed analysis of the impact of U. S. energy policy on the sectors most directly affected by policy changes. The nine sectors included in the model are:

(1) agriculture, nonfuel mining, and construction
(2) manufacturing, excluding petroleum refining
(3) transportation
(4) communications, trade, and services
(5) coal mining
(6) crude petroleum and natural gas
(7) petroleum refining
(8) electric utilities
(9) gas utilities.

Our interindustry model includes a model of demand for inputs and supply of output for each of the nine industrial sectors. The model is closed by balance equations between demand and supply for the products of each of the nine sectors.

The principal innovation of our interindustry model is that the input-output coefficients are treated as endogenous variables rather than exogenously given parameters. Our model for producer behavior determines the input-output coefficients for each of the nine sectors listed above as functions of the prices of products of all sectors, the prices of labor and capital services, and the prices of competing imports. We determine the prices of all nine products and the matrix of input-output coefficients simultaneously. In conventional input-output analysis the technology of each sector is taken as fixed at any point of time. Prices are determined as functions of the input-output coefficients, but the input-output coefficients themselves are treated as exogenously given parameters. Our approach integrates conventional input-output analysis with a deter-

---

A recent compendium of research on input-output analysis is Carter and Brody [10].

[3] A more detailed presentation of our approach is contained in Jorgenson, Berndt, Christensen, and Hudson [26].

mination of the structure of technology through models of supply for each industrial sector.

The second component of our framework for energy policy analysis is a macro-econometric growth model. The complete model consists of endogenous business and household sectors and exogenous foreign and government sectors. The chief novelty of our growth model is the integration of demand and supply conditions for consumption, investment, capital, and labor. The model is made dynamic by links between investment and changes in capital stock and between capital service prices and changes in investment goods prices. The model determines both components of gross national product in real terms, as generated by conventional macro-econometric models, and relative prices of labor and capital services required by our econometric model of interindustry transactions.

Our approach to the analysis of macro-economic activity can be contrasted with the analysis that underlies macro-econometric models used for short-term forecasting. Short-term forecasting is based on the projection of demand by foreign and government sectors and the determination of the responses of households and businesses in the form of demands for consumption and investment goods. The underlying economic theory is essentially the Keynesian multiplier, made dynamic by introducing lags in the responses of households and businesses to changes in income. In short-term macro-econometric models the supply side is frequently absent or present in only rudimentary form.[4] Our approach integrates the determinants of demand employed in conventional macro-econometric models with the determinants of supply; this integration is essential to the successful implementation of our interindustry model for longer run analyses.

Given a framework that incorporates the determinants of demand and supply for energy in the U. S. economy, our first objective is to provide a reference point for the analysis of energy policy by establishing detailed projections of demand and supply, price and cost, and imports and exports for each of the nine industrial sectors included in our model. For this purpose we project the level of activity in each industrial sector and relative prices for the products of all sectors for the years 1975 through 2000. Our projections include the level of macro-economic activity in the U. S. economy and the matrix of input-output coefficients for each year. Projections for the five industrial sectors that form the energy sector of the U. S. economy provide the basis for translating our detailed projections into the energy balance framework that has become conventional in the analysis of patterns of energy utilization.[5]

Our interindustry approach imposes the same consistency requirements as the energy balance approach, namely, that demand is equal to supply in physical terms for each type of energy. In addition, our approach requires that demand and supply be consistent with the same structure of energy prices. This additional consistency requirement is absent from energy balance projections and requires the integration of energy balance projections with projections of energy prices. Our interindustry model provides a means of combining these projections within a framework that also includes prices and interindustry transactions for the sectors that consume but do not produce energy.

---

[4] In the Klein-Goldberger model the determination of prices can be completely suppressed with a resulting improvement in forecasting accuracy for real magnitudes. See Suits [38] and Goldberger [22].

[5] The energy balance framework has been employed by Dupree and West [20] and the National Petroleum Council [35, 36].

To illustrate the application of our model to the analysis of U. S. energy policy we have analyzed the effects of tax policies to stimulate energy conservation on the future pattern of energy utilization. Our methodology for policy analysis begins with a set of projections that assume no major new departures in energy policy. We then prepare an alternative set of projections incorporating the proposed change in policy. In analyzing the impact of tax policy we have incorporated the effect of energy taxes on demand and supply for energy. We find that price increases provide the economic incentive for the adoption of energy conservation measures that will result in considerable savings of energy. Tax policies or other measures to increase the price of energy could result in U. S. independence from energy imports by 1985.

We present our model for interindustry transactions in Section 2 of the paper. In Section 3 we present econometric models of producer behavior for each of the nine sectors included in our interindustry model. We then outline our macro-econometric growth model in Section 4. In Section 5 we present projections of economic activity and energy utilization for the period 1975 through 2000. In Section 6 we discuss a tax program for stimulating energy conservation and eliminating reliance of the U. S. economy on energy imports.

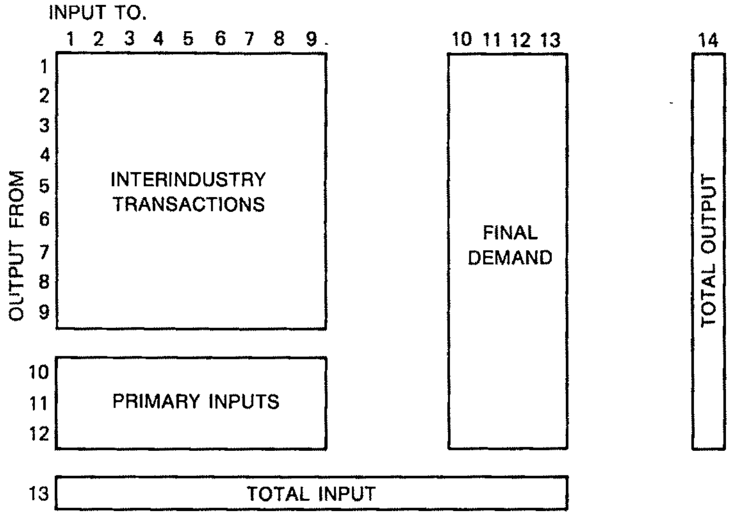
2. Interindustry model

■ The first component of our framework for energy policy analysis is a model of interindustry transactions for the United States. Rather than analyzing energy utilization in isolation, we begin with an analysis of the entire U. S. economy and then proceed to a detailed examination of the energy sector as one among many interdependent components of the economy. This perspective is necessarily more complex and more detailed than traditional perspectives on the analysis of the energy sector, but is indispensable to the study of the interaction of energy resources and the growth of the U. S. economy.

Our interindustry model permits the analysis of the entire chain of production from the purchase of primary inputs through the various intermediate stages of production to the emergence of final products to be absorbed in consumption, investment, government, or export final demand. The structure of production includes all of U. S. domestic supply of goods and services, but our specification provides for detailed analysis of the impact of U. S. energy policy on the sectors most directly affected by policy changes. We have classified production into nine industrial sectors, each of which purchases primary inputs, makes purchases from and sales to the other producing sectors, and sells finished output to final users. The flow of interindustry transactions is represented in diagrammatic form in Figure 1.

FIGURE 1

INTERINDUSTRY TRANSACTIONS. DIAGRAMMATIC REPRESENTATION



INTERMEDIATE SECTORS:

1 AGRICULTURE, NONFUEL MINING AND CONSTRUCTION.
2 MANUFACTURING, EXCLUDING PETROLEUM REFINING
3. TRANSPORT.
4. COMMUNICATIONS, TRADE, SERVICES.
5 COAL MINING.
6. CRUDE PETROLEUM AND NATURAL GAS.
7. PETROLEUM REFINING
8. ELECTRIC UTILITIES.
9 GAS UTILITIES

PRIMARY INPUTS, ROWS

10 IMPORTS
11 CAPITAL SERVICES.
12 LABOR SERVICES

FINAL DEMAND, COLUMNS:

10 PERSONAL CONSUMPTION EXPENDITURES.
11. GROSS DOMESTIC PRIVATE INVESTMENT.
12. GOVERNMENT PURCHASES OF GOODS AND SERVICES.
13. EXPORTS

Our interindustry model consists of balance equations between supply and demand for the products of each of the nine sectors included in the model. The model also includes accounting identities between the value of domestic availability of these products and the sum of values of intermediate input into each industry, value added in the industry, and imports of competing products. Demands for the products include demands for use as inputs by each of the nine sectors included in the model. The rest of domestic availability is allocated among four categories of final demand: personal consumption expenditures, gross private domestic investment, government expenditures, and exports.

In the model for projecting energy demand and supply we take the levels of final demand for all industries from the macro-econometric model presented in Section 4, below. Second, for the five energy sectors of the model we take the price and quantity of imports to be exogenous. For the four nonenergy sectors we take the prices of imports as exogenous and determine import quantities along with the quantities of capital and labor services in each industry.[6] The prices of capital and labor services are determined within the macro-econometric model. We take the quantities of exports and government purchases of the output of each industry as exogenous. We also take the allocation of investment among the industries of origin to be exogenous.

Our interindustry model consists of models of producer behavior for each of the nine industries included in the model. Producer behavior in each industry can be characterized by input-output coefficients for the input of products of each of the nine sectors, inputs of capital and labor services, and, for the four nonenergy sectors, the level of competitive imports. An interindustry approach to the study of energy resources is essential, since most energy is consumed as an intermediate rather than a final product of the economy. Examples of intermediate products would be fossil fuels consumed by the electric generating sector. Examples of final products would be gasoline and heating oil consumed by the household and government sectors. Energy balance models provide projections of the levels of both intermediate and final demand.

Given the prices of domestic availability of the output of each sector included in our model, we determine the allocation of personal consumption expenditures among commodity groups distinguished in the model, using our model of consumer behavior. Personal consumption expenditures include deliveries to the household sector by eight of the nine sectors included in our interindustry model of the producing sector. There are no direct deliveries of crude petroleum and natural gas to personal consumption expenditures. These products are delivered first to the petroleum refining and gas utility sectors and then to personal consumption expenditures and to other categories of intermediate and final demand.

Personal consumption expenditures also include noncompetitive imports and the services of dwellings and consumers' durables. The levels of personal consumption expenditures on each of the eleven commodity groups included in our model of the household sector are determined from the projected level of personal consumption expenditures from the macro-econometric model, from the prices of domestic availability of the output of each sector included in the interindustry model, and from the prices of noncompetitive imports, consumers' durables services, and housing services. The price of noncompetitive imports is taken to be exogenous. The capital service prices for consumers' durables services and housing services are determined from the price of capital services determined in the macro-econometric model.

The equations representing the balance of demand and supply for each of the nine sectors of the interindustry model set domestic availability equal to the sum of intermediate demands and final demand. Intermediate demands are determined simultaneously with the levels of output of each industry, given input-output coefficients determined in the model of producer behavior. The input-output coefficients are determined simultaneously with the prices of domestic availability of the output of each industry. Finally, levels of capital and labor services for all sectors and competitive imports for the four nonenergy sectors are determined from the levels of domestic availability and the corresponding input-output coefficients. These levels can be compared with the levels projected in the macro-econometric model.

---

[6] Energy imports are significant only for crude and refined petroleum products and natural gas. For the period 1958 to 1972 petroleum imports were subject to a system of quotas. Natural gas imports are subject to regulation by the Federal Power Commission. For a discussion of the petroleum import quota system, see Burrows and Domencich [8].

☐ **Interindustry transactions.** We first describe our model of interindustry transactions and then outline the application of this model to the projection of energy demand and supply. Our notation is as follows:

XIJ = intermediate demand for the output of industry I by industry J;
YI = final demand for output of industry I;
XI = domestic availability of the output of industry I;
PI = price of the output of industry I.

To simplify the notation we take the price of the output of each industry to be the same in all uses. The deflators for each category of intermediate and final demand can differ. In projecting energy demand and supply we take the ratios of the deflator for the individual categories of demand to the deflator for domestic availability of output of the industry to be exogenous.

The interindustry model consists of equality between demand and supply for each of the nine sectors included in the model. The balance equations for the nine sectors are:

$$XI = \sum_{J=1}^{9} XIJ + YI, \qquad (I = 1, 2 \ldots 9). \tag{1}$$

In addition, the model includes accounting identities between the value of domestic availability and the sum of values of intermediate input into the industry, value added in the industry, and, for the four-nonenergy sectors, the imports of competing products:

$$PI^{*}XI = \sum_{J=1}^{9} PJ^{*}XJI + PK^{*}KI + PL^{*}LI + PRI^{*}RI, \qquad (I = 1, 2 \ldots 9), \tag{2}$$

where:

KI = quantity of capital services in industry I;
LI = quantity of labor services in industry I;
RI = competitive imports of the output of industry I;
PK = price of capital services;
PL = price of labor services;
PRI = price of competitive imports to industry I.

Again, prices of capital and labor services can differ among industries. To simplify notation we take the prices of these productive factors to be the same in all industries. In projecting energy demand and supply we take the ratios of service prices for each industry to the corresponding prices from the macro-econometric model to be exogenous.

Our interindustry model includes models of producer behavior for each of the nine industrial sectors included in the model. These models of producer behavior can be derived from price possibility frontiers for the nine sectors:

$$AI^{*}PI = GI(P1, P2, \ldots P9; PK, PL, PRI), \qquad (3)$$
$$(\mathrm{I} = 1, 2 \ldots 9),$$

where $AI (I = 1, 2 \ldots 9)$ is an index of the level of Hicks-neutral technology in industry I. The price possibility frontier for each sector can be derived from price possibility frontiers for each of the three submodels employed in our analysis of production structure:[7]

(1) a model giving the price of output as a function of prices of four aggregate inputs in each sector—capital (K), labor (L), energy (E), and materials (M);

(2) a model giving the price of aggregate energy input in each sector as a function of the prices of the five types of energy included in the model—coal, crude petroleum and natural gas, refined petroleum products, electricity, and gas as a product of gas utilities;

(3) a model giving the price of aggregate nonenergy input in each sector as a function of the prices of the five types of nonenergy input into each sector—agriculture, manufacturing, transportation, communications, and, for the four nonenergy sectors, competitive imports.

Given the prices of capital services, labor services, and competitive imports in each of the four nonenergy sectors, we can determine the prices of domestic availability of output PI ($I = 1, 2 \ldots 9$) for all nine sectors. To determine these prices we solve twenty-seven equations for prices of domestic availability, prices of aggregate energy input, and prices of aggregate nonenergy input into all nine sectors. This system of twenty-seven equations consists of three equations for each sector. These three equations correspond to production possibility frontiers for each of the three submodels for each sector. In these computations we are making use of a nonsubstitution theorem of the type first discussed by Samuelson.[8] This theorem states that for given prices of the factors of production and competitive imports, the prices of domestic availability of the output of each sector are independent of the composition of final demand.

The second step in our analysis of interindustry transactions is to derive input-output coefficients for each of the nine industrial sectors included in our interindustry model. The input-output coefficients can be expressed as functions of the prices. First, the relative share of the jth intermediate input can be determined from the identity:

$$\frac{\partial \ln PI}{\partial \ln PJ} = \frac{PJ^{*}XJI}{PI^{*}XI} = \frac{PJ}{PI}{}^{*}AJI, \quad (I, J = 1, 2 \ldots 9), \qquad (4)$$

where AJI is the input-output coefficient corresponding to XJI; it represents the input of the output of industry J per unit of output of industry *I*. Similar identities determine the relative shares of capital and labor services and competitive imports.[9]

---

[7] For a detailed interpretation of the price possibility frontier, see Christensen, Jorgenson, and Lau [18], especially pp. 32–33.

[8] See Samuelson [37].

Second, we can divide the relative shares by the ratio of the price of domestic availability of the output of the *J*th industry, *PJ* to the price for the *I*th industry, *PI*, to obtain the input-output coefficients:

$$\frac{XJI}{XI} = AJI(P1, P2 \ldots P9; PK, PL, PRI), \qquad (5)$$
$$(I, J = 1, 2 \ldots 9);$$

and

$$\frac{KI}{XI} = AKI(P1, P2 \ldots P9; PK, PL, PRI),$$
$$\frac{LI}{XI} = ALI(P1, P2 \ldots P9; PK, PL, PRI), \qquad (6)$$
$$\frac{RI}{XI} = ARI(P1, P2 \ldots P9; PK, PL, PRI),$$
$$(I = 1, 2 \ldots 9).$$

For each industry we derive the input-output coefficients in two steps. First, we determine the input-output coefficients for the aggregate inputs—capital (K), labor (L), energy (E) and materials (M). Second, we determine the input-output coefficients for the input of each type of energy input per unit of total energy input and the input of each type of nonenergy input per unit of total nonenergy input. To obtain the input-output coefficients required for our interindustry model we multiply the input-output coefficients for each type of energy by the input-output coefficient for total energy. Similarly, we multiply the input-output coefficients for each type of nonenergy input by the input-output coefficient for total nonenergy input. We obtain input-output coefficients for capital services, labor services, five types of energy inputs into each sector, and five types of nonenergy inputs into each sector.

The input-output coefficients for each of the nine industrial sectors included in our model of interindustry transactions are functions of the prices of capital services, labor services, and competitive imports for the four nonenergy sectors and the prices of domestic availability of the output of each of the nine sectors. The prices of domestic availability are functions of the prices of capital services, labor services, and competitive imports for the four nonenergy sectors. By the nonsubstitution theorem both prices of domestic availability and input-output coefficients are independent of the composition of final demand.[10]

---

[9] For further discussion of the model of producer behavior, see Section 3, below.

[10] The idea of treating input-output coefficients as functions of prices can be traced to Walras [41], especially pp. 382–392; this approach has been extensively discussed by Samuelson [37], pp. 513–536, and Morishima [33], pp. 54–92. A more influential idea is to model trends in input-output coefficients without treating them as part of a model of producer behavior. This alternative approach has been employed by Leontief [31], Carter [9] and Almon, *et al.* [1]. Comparisons of input-output coefficients for 1947, 1958, and 1961 are given by Carter [9] and Vaccara [40].

☐ **Final demand.** Final demand for domestic availability of the output of each of the nine sectors included in our interindustry model is allocated among personal consumption expenditures, gross private domestic investment, government expenditures, and exports. In projecting energy demand and supply we take aggregate levels of each category of final demand from our macro-economic projections. We allocate personal consumption expenditures among the nine sectors included in our model, employing aggregate personal consumption expenditures as total expenditures, on the basis of the prices of domestic availability of the output of all nine sectors. Government expenditures and exports of the output of each sector are exogenous. Imports of the output of the five energy sectors are also exogenous so that we include only exports net of imports in final demand for these sectors. We take aggregate private domestic investment from our macro-economic projections. We take the relative proportion of investment in the output of each industrial sector included in our interindustry model to be exogenous.

The final step in determining the level and composition of interindustry transactions is to determine the levels of output, employment, and utilization of capital for each of the nine industrial sectors included in our model and competitive imports for the four nonenergy sectors included in the model. This part of our model coincides with conventional input-output analysis. Given the input-output coefficients for all nine sectors, we can determine the level of output for each sector for any given levels of final demand for the output of all nine sectors. We present projected matrices of interindustry transactions in energy for the years 1975, 1980, 1985, and 2000 in Section 5 below. We also present projections of energy prices for each of these years.

Final demand for domestic availability of the output of each of the nine sectors included in the model is allocated among consumption, investment, government expenditures, and exports:

$$YI = CI + II + GI + ZI, \qquad (I = 1, 2 \ldots 9), \qquad (7)$$

where:

CI = personal consumption expenditures on the output of industry I;
II = gross private domestic investment in the output of industry I (the sum of gross private fixed investment and net inventory change);
GI = government expenditure on the output of industry I;
ZI = exports of the output of industry I (exports less imports for the five energy sectors).

In our model for projecting energy demand and supply we take the levels of final demand for all industries from the macroeconomic projections. We link our interindustry model to our macro-econometric model through the identities:

$$PC^{*}C = \sum_{I=1}^{9} PI^{*}CI,$$

$$PI^*I = \sum_{I=1}^{9} PI^*II, \qquad (8)$$

$$PG^*G = \sum_{I=1}^{9} PI^*GI,$$

$$PZ^*Z = \sum_{I=1}^{9} PI^*ZI.$$

The values of personal consumption expenditures PC^*C, gross private domestic investment PI^*I, government expenditure PG^*G, and exports PZ^*Z from the macro-econometric model are set equal to the sums of each of these categories of expenditures over all nine industries included in the interindustry model.

In the macro-econometric model government expenditures on goods and services are divided into two parts:

$$PG^*G = PIG^*IG + PCG^*CG, \qquad (9)$$

where

IG = quantity of government expenditures on investment goods;
CG = quantity of government expenditures on consumption goods;
PIG = price of government expenditures on investment goods;
PCG = price of government expenditures on consumption goods.

In making projections of energy demand and supply we project total government expenditures in current and constant prices. Government expenditures on goods and services are exogenous to the macro-econometric model. We then allocate total government expenditures among the nine industry groups included in the interindustry model.

In our macro-econometric model net exports of goods and services are taken to be exogenous. For the purposes of projecting energy demand and supply we divide net exports between imports and exports, allocate exports among the nine industry groups included in the model, and project the prices of competitive imports for each of the nine sectors of the model. Since net exports in current and constant prices are exogenous to the macro-econometric model, this change in the treatment of net exports does not alter the structure of the complete model. In projecting energy demand and supply we take the prices of imports for each industrial sector together with levels of capital and labor services for each sector.

We project gross private domestic investment in current and constant prices in our macro-econometric model. To project energy demand and supply we allocate gross private domestic investment among the nine industry groups included in the model. The relative proportions of investment originating in each sector are taken to be exogenous. In a completely dynamic model the allocation of investment by sector of origin and sector of destination would be endogenous. Our macro-econometric model incorporates the dynamics of saving and investment only in the projection of total investment. The allocation of capital by sector of destination is endogenously determined, but the allocation of investment by sector of origin is exogenous.

The final step in determining final demand for the domestic availability of the output of each sector included in our interindustry model is to allocate personal consumption expenditures among the products of the nine sectors included in the model and expenditures on noncompetitive imports and the services of consumers' durables, which are not included among the products of the nine sectors. For this purpose we employ an econometric model of consumer behavior. This model is based on an indirect utility function that can be represented in the form:[11]

$$\ln V = \ln \dot{V}\left(\frac{P1}{PC^*C}, \frac{P2}{PC^*C} \; \cdots \frac{P11}{PC^*C} \right), \qquad (10)$$

where V is the level of utility, PI is the price of the Ith commodity, and PC^*C is total personal consumption expenditures. There are eleven commodity groups included in our model of consumer behavior: one for each of the nine industrial sectors, excluding crude petroleum and natural gas, and one each for housing, services of consumers' durables, and noncompetitive imports.

For each commodity group we take the budget share to be fixed. This assumption corresponds to a linear logarithmic indirect utility function. The quantity demanded for each of the eleven commodity groups included in our model of consumer behavior is a function of the prices of the corresponding commodity group and the level of total personal consumption expenditures. We determine the level of total personal consumption expenditures in our macroeconometric model. We determine the prices of domestic availability of the output of each sector in the interindustry model from our models of producer behavior. Given total expenditures and the prices, we can determine the quantities demanded of the output of each sector of the model for personal consumption expenditures.

Final demand for all nine industrial sectors included in our interindustry model is the sum of final demands for personal consumption expenditures, gross private domestic investment, government expenditures, and exports. We add personal consumption expenditures for the output of communications, trade, and services, less housing, to personal consumption expenditures for housing to obtain personal consumption expenditures on communications, trade, and services. Otherwise, there is a direct correspondence between commodity groups in our model of consumer behavior and the sectors included in our interindustry model. Given all final demands, the prices of domestic availability of the output of each sector and the matrix of input-output coefficients, we can determine the matrix of interindustry transactions in both current and constant prices.

The equations representing the balance of demand and supply for each of the nine industrial sectors included in our interindustry model can be presented in the form:

---

[11] For a detailed discussion of the indirect utility function, see Christensen, Jorgenson, and Lau [19].

$$\begin{aligned} XJ &= \sum_{J=1}^{9} XIJ + YI, \\ &= \sum_{J=1}^{9} AIJ^{*}XJ + CI + II + GI + ZI, \quad (I = 1, 2 \ldots 9). \end{aligned} \qquad (11)$$

The input-output coefficients $\{AIJ\}$ are determined together with the prices of domestic availability of the output of each industry $\{PI\}$. Given prices and the level of aggregate personal consumption expenditures, the levels of personal consumption expenditures $\{CI\}$ are determined. The remaining components of final demand $\{II, GI, ZI\}$ are projected by industry of origin. These projections are consistent with levels of gross private domestic investment, government expenditures and exports from our macro-econometric model.

In matrix form the demand and supply balance equation for the model can be represented as:

$$x = Ax + y, \qquad (12)$$

where x and y are vectors of outputs and final demands:

$$x = \begin{vmatrix} X1 \\ X2 \\ \cdot \\ \cdot \\ \cdot \\ X9 \end{vmatrix}, \qquad y = \begin{vmatrix} Y1 \\ Y2 \\ \cdot \\ \cdot \\ \cdot \\ Y9 \end{vmatrix},$$

and A is the matrix of input-output coefficients:

$$A = \begin{vmatrix} A11 & A12 & \ldots & A19 \\ A21 & A22 & \ldots & A29 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ A91 & A92 & \ldots & A99 \end{vmatrix}.$$

Levels of domestic availability of the output of each vector are obtained by solving this system of equations:

$$x = (I - A)^{-1}y. \qquad (13)$$

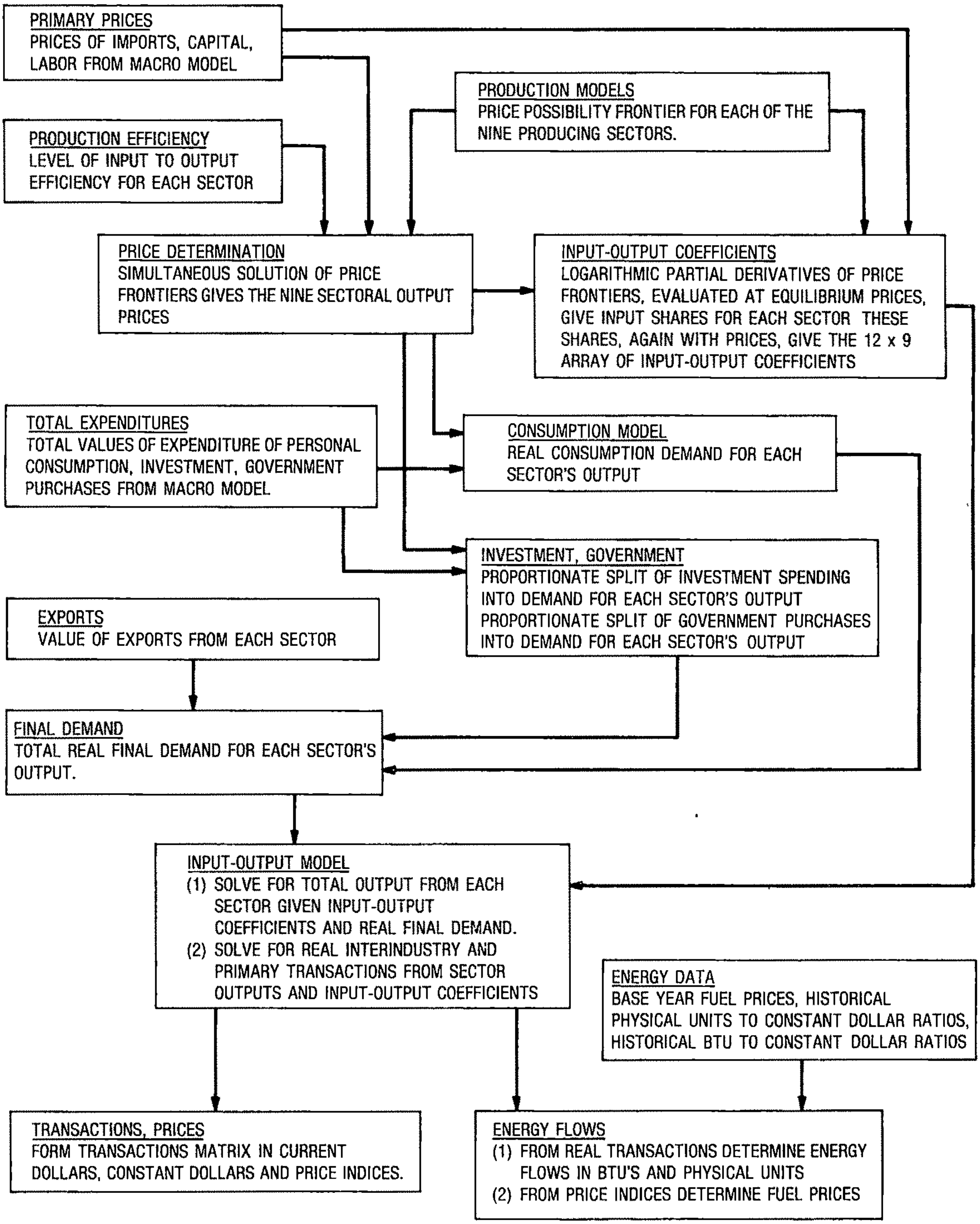
Levels of capital and labor services and competitive imports are determined from the levels of domestic availability and the corresponding input-output coefficients:

$$\begin{aligned} KI &= AKI^{*}XI, \\ LI &= ALI^{*}XI, \\ RI &= ARI^{*}XI, \\ (I &= 1, 2 \ldots 9). \end{aligned} \qquad (14)$$

The input-output coefficients for capital and labor services and competitive imports are functions of the prices of the outputs of the nine sectors included in our interindustry model, the prices of capital and labor services and the prices of competitive imports for the four nonenergy sectors of the model. Our complete econometric model for interindustry transactions is presented in diagrammatic form in Figure 2.

FIGURE 2

INTERINDUSTRY ECONOMETRIC MODEL DIAGRAMMATIC REPRESENTATION


PRIMARY PRICES
PRICES OF IMPORTS, CAPITAL, LABOR FROM MACRO MODEL
PRODUCTION MODELS
PRICE POSSIBILITY FRONTIER FOR EACH OF THE NINE PRODUCING SECTORS.
PRODUCTION EFFICIENCY
LEVEL OF INPUT TO OUTPUT EFFICIENCY FOR EACH SECTOR
PRICE DETERMINATION
SIMULTANEOUS SOLUTION OF PRICE FRONTIERS GIVES THE NINE SECTORAL OUTPUT PRICES
INPUT-OUTPUT COEFFICIENTS
LOGARITHMIC PARTIAL DERIVATIVES OF PRICE FRONTIERS, EVALUATED AT EQUILIBRIUM PRICES, GIVE INPUT SHARES FOR EACH SECTOR THESE SHARES, AGAIN WITH PRICES, GIVE THE 12 x 9 ARRAY OF INPUT-OUTPUT COEFFICIENTS
TOTAL EXPENDITURES
TOTAL VALUES OF EXPENDITURE OF PERSONAL CONSUMPTION, INVESTMENT, GOVERNMENT PURCHASES FROM MACRO MODEL
CONSUMPTION MODEL
REAL CONSUMPTION DEMAND FOR EACH SECTOR'S OUTPUT
INVESTMENT, GOVERNMENT
PROPORTIONATE SPLIT OF INVESTMENT SPENDING INTO DEMAND FOR EACH SECTOR'S OUTPUT
PROPORTIONATE SPLIT OF GOVERNMENT PURCHASES INTO DEMAND FOR EACH SECTOR'S OUTPUT
EXPORTS
VALUE OF EXPORTS FROM EACH SECTOR
FINAL DEMAND
TOTAL REAL FINAL DEMAND FOR EACH SECTOR'S OUTPUT.
INPUT-OUTPUT MODEL
(1) SOLVE FOR TOTAL OUTPUT FROM EACH SECTOR GIVEN INPUT-OUTPUT COEFFICIENTS AND REAL FINAL DEMAND.
(2) SOLVE FOR REAL INTERINDUSTRY AND PRIMARY TRANSACTIONS FROM SECTOR OUTPUTS AND INPUT-OUTPUT COEFFICIENTS
ENERGY DATA
BASE YEAR FUEL PRICES, HISTORICAL PHYSICAL UNITS TO CONSTANT DOLLAR RATIOS, HISTORICAL BTU TO CONSTANT DOLLAR RATIOS
TRANSACTIONS, PRICES
FORM TRANSACTIONS MATRIX IN CURRENT DOLLARS, CONSTANT DOLLARS AND PRICE INDICES.
ENERGY FLOWS
(1) FROM REAL TRANSACTIONS DETERMINE ENERGY FLOWS IN BTU'S AND PHYSICAL UNITS
(2) FROM PRICE INDICES DETERMINE FUEL PRICES

3. Producer behavior

■ Our interindustry model includes econometric models of producer behavior for each of the nine industrial sectors included in the model.[12] In implementing an econometric model of producer behavior for each sector our primary objective is to explore the

---

[12] This Section is based on Berndt and Jorgenson [5].

interrelationships between relative demand for energy and relative demand for capital services, labor services, and nonenergy inputs. Similarly, we wish to explore the interrelationships among relative demands for the five types of energy included in our model—coal, crude petroleum and natural gas, refined petroleum products, electricity, and gas as a product of gas utilities. We have imposed a structure on the price possibility frontier that permits us to deal with relative demand for energy as a whole and relative demands for the five types of energy included in our model as two separate problems.

To construct a model of the interrelationship of relative demands for energy, capital services, labor services, and nonenergy inputs, we first define groups of inputs that are aggregates of the twelve inputs included in our model of interindustry production structure. These commodity groups are:

(1) Capital (K).
(2) Labor (L).
(3) Energy (E). This group consists of inputs of coal, crude petroleum and natural gas, refined petroleum products, electricity, and gas as a product of gas utilities.
(4) Materials (M). This group consists of inputs of agriculture, manufacturing, transportation, communications, trade, and services, and competitive imports for the nonenergy sectors.

We first construct a model for producer behavior in terms of the four aggregates—capital, labor, energy, and materials. We represent the price of domestic availability of the output of each sector as a function of the prices of each of the aggregates. A sufficient condition for the price possibility frontier to be defined on the prices of the four aggregates is that the overall price possibility frontier is separable and homogeneous in the inputs within each aggregate.[13] The price possibility frontier is separable in the commodities within an aggregate if and only if the ratio of the relative shares of any two commodities within an aggregate is independent of the prices of commodities outside the aggregate.[14] For example, the five types of energy make up an appropriate aggregate if the relative value shares of any two types of energy depend only on the prices of energy and not on the prices of nonenergy intermediate inputs or the prices of capital and labor services.

The second step in constructing a model of producer behavior is to represent the price possibility frontier for the energy and materials aggregates as functions of the prices of inputs that make up each of the aggregates. For the energy aggregate the price of energy is represented as a function of the prices of the five types of energy that make up the aggregate—coal, crude petroleum and natural gas, refined petroleum products, electricity, and gas as a product of gas utilities. For the materials aggregate the price of materials is represented as a function of the five types of inputs that make up the aggregate—agriculture, manufacturing, transportation, communications, trade, and services, and competitive imports for the nonenergy sectors.

---

[13] See Christensen, Jorgenson, and Lau [18], pp. 29–32.
[14] See Leontief [29].

☐ **Econometric specification.** The system of relative demand functions employed in our econometric model of producer behavior for each of the nine industrial sectors of our model is generated from the price possibility frontier for the corresponding sector. For each of the three submodels that make up our model of producer behavior we represent the price possibility frontier by a function that is quadratic in the logarithms of the prices of the inputs into the sector. The resulting price possibility frontier provides a local second-order approximation to any price possibility frontier. We refer to our representation as the transcendental logarithmic price possibility frontier or, more simply, the *translog price possibility frontier*. The price possibility frontier is a transcendental function of the logarithms of the prices of inputs. The translog price possibility frontier was introduced by Christensen, Jorgenson, and Lau.[15]

As an example, the price possibility frontier for the aggregate (*KLEM*) submodel takes the form:

$$\ln AI + \ln PI = \alpha^I_0 + \alpha^I_K \ln PK + \alpha^I_L \ln PL + \alpha^I_E \ln PE + \alpha^I_M \ln PM + \frac{1}{2}\left[\beta^I_{KK}(\ln PK)^2 + \beta^I_{KL} \ln PK \ln PL + \ldots\right], \quad (15)$$

where *PK* is the price of capital services, *PL* the price of labor services, *PE* the price of energy, and *PM* the price of materials. For this form of the price possibility frontier, the equations for the relative shares of the four input aggregates take the form:

$$\begin{aligned}
\frac{PK^*KI}{PI^*XI} &= \alpha^I_K + \beta^I_{KK}\ln PK + \beta^I_{KL}\ln PL + \beta^I_{KE}\ln PE + \beta^I_{KM}\ln PM,\\
\frac{PL^*LI}{PI^*XI} &= \alpha^I_L + \beta^I_{LK}\ln PK + \beta^I_{LL}\ln PL + \beta^I_{LE}\ln PE + \beta^I_{LM}\ln PM,\\
\frac{PE^*EI}{PI^*XI} &= \alpha^I_E + \beta^I_{EK}\ln PK + \beta^I_{EL}\ln PL + \beta^I_{EE}\ln PE + \beta^I_{EM}\ln PM, \qquad (16)\\
\frac{PM^*MI}{PI^*XI} &= \alpha^I_M + \beta^I_{MK}\ln PK + \beta^I_{ML}\ln PL + \beta^I_{ME}\ln PE + \beta^I_{MM}\ln PM,\\
&\qquad (I = 1, 2 \ldots 9),
\end{aligned}$$

where *KI* is the quantity of capital services in the *I*th sector, *LI* the quantity of labor services, *EI* the quantity of energy input, and *MI* the quantity of materials input.[16]

The dependent variable in each of the four functions generated from the translog price possibility frontier is the relative share of the corresponding input. To derive the input-output coefficient for that input, we divide the relative share by the ratio of the price of the input to the price of the output of the sector. For example, the input-output coefficients for capital services are:

---

[15] See Christensen, Jorgenson, and Lau [17].

[16] A *KLEM* model for total U. S. manufacturing based on the translog price possibility frontier has been developed by Berndt and Wood [6]. Berndt and Christensen [2, 3, 4] have developed models of capital-labor substitution for U. S. manufacturing based on the translog production function, which is dual to the translog price possibility frontier.

$$AKI = \frac{KI}{XI} = (\alpha^I_K + \beta^I_{KK} \ln PK + \beta^I_{KL} \ln PL$$
$$+ \beta^I_{KE} \ln PE + \beta^I_{KM} \ln PM)/(PK/PI), \; (I = 1, 2 \ldots 9).$$

Similar expressions can be obtained for the input-output coefficients for labor services, energy, and materials.

The value of domestic availability of the output of each sector is equal to the sum of the values of capital and labor services in that sector and the value of energy and nonenergy inputs into the sector:

$$PI^*XI = PK^*KI + PL^*LI + PE^*EI + PM^*MI, \quad (I = 1, 2 \ldots 9). \tag{17}$$

Given this accounting identity, the relative shares of the four aggregate inputs into each sector add to unity. The parameters of the four relative demand functions for capital and labor services and energy and nonenergy inputs must satisfy the restrictions:

$$\begin{aligned} &\alpha^I_K + \alpha^I_L + \alpha^I_E + \alpha^I_M = 1, \\ &\beta^I_{KK} + \beta^I_{LK} + \beta^I_{EK} + \beta^I_{MK} = 0, \\ &\beta^I_{KL} + \beta^I_{LL} + \beta^I_{EL} + \beta^I_{ML} = 0, \\ &\beta^I_{KE} + \beta^I_{LE} + \beta^I_{EE} + \beta^I_{ME} = 0, \\ &\beta^I_{KM} + \beta^I_{LM} + \beta^I_{EM} + \beta^I_{MM} = 0, \\ &\qquad (I = 1, 2 \ldots 9). \end{aligned} \tag{18}$$

Given estimates of the parameters of any three equations for the relative shares, estimates of the parameters of the fourth equation can be determined from these restrictions.

The logarithm of the price possibility frontier for each sector is twice differentiable in the logarithms of the prices of inputs, so that the Hessian of this function is symmetric. This gives rise to a set of restrictions relating the parameters of cross partial derivatives. For the aggregate (*KLEM*) submodel three of these restrictions are explicit in the three equations we estimate directly, namely:

$$\begin{aligned} &\beta^I_{KL} = \beta^I_{LK}, \\ &\beta^I_{KE} = \beta^I_{EK}, \\ &\beta^I_{LE} = \beta^I_{EL}, \\ &\qquad (I = 1, 2 \ldots 9). \end{aligned} \tag{19}$$

In addition, we estimate the parameters β^I_{MK}, β^I_{ML}, and β^I_{ME} $(I = 1, 2 \ldots 9)$ from the equations:

$$\begin{aligned} &\beta^I_{MK} = -\beta^I_{KK} - \beta^I_{LK} - \beta^I_{EK}, \\ &\beta^I_{ML} = -\beta^I_{KL} - \beta^I_{LL} - \beta^I_{EL}, \\ &\beta^I_{ME} = -\beta^I_{KE} - \beta^I_{LE} - \beta^I_{EE}, \\ &\qquad (I = 1, 2, \ldots 9), \end{aligned}$$

so that three additional symmetry restrictions are implicit in the equations we estimate, namely:

$$\begin{aligned} &\beta^I_{KM} = \beta^I_{MK}, \\ &\beta^I_{LM} = \beta^I_{ML}, \\ &\beta^I_{EM} = \beta^I_{ME}, \\ &\qquad (I = 1, 2 \ldots 9). \end{aligned} \tag{20}$$

For each of the nine industrial sectors, the aggregate (*KLEM*) submodel involves six symmetry restrictions.

The price possibility frontier for each sector is homogeneous of degree one; proportional changes in the prices of all inputs result in a proportional change in the price of output. Homogeneity of the price possibility frontier is implied by the symmetry restrictions outlined above and the restrictions implied by the accounting identity between the value of output and the value of input. In the absence of the symmetry restrictions and the restrictions implied by the accounting identity between the value of output and the value of input, the aggregate (*KLEM*) submodel involves twenty unknown parameters. Taking these restrictions into account, we reduce the number of unknown parameters to nine.

We have presented the aggregate (*KLEM*) submodel of our model of producer behavior in detail. The forms of the energy (*E*) and materials (*M*) submodels are analogous to the form of the aggregate submodel. For the energy submodel we can write the translog price possibility frontier in the form:

$$\begin{aligned}\ln PE = \alpha_0^{EI} + \alpha_1^{EI} \ln PE1 + \alpha_2^{EI} \ln PE2 + \alpha_3^{EI} \ln PE3 \\ + \alpha_4^{EI} \ln PE4 + \alpha_5^{EI} \ln PE5 \qquad (21) \\ + \frac{1}{2} [\beta_{11}^{EI} (\ln PE1)^2 + \beta_{12}^{EI} \ln PE1 \ln PE2 + \ldots], \\ (I = 1, 2 \ldots 9),\end{aligned}$$

where *PE*1 is the price of coal, *PE*2 the price of crude petroleum and natural gas, *PE*3 the price of refined petroleum products, *PE*4 the price of electricity, and *PE*5 the price of gas as a product of gas utilities. Similarly, we can write the translog price possibility frontier for the materials submodel in the form:

$$\begin{aligned}\ln PM = \alpha_0^{MI} + \alpha_1^{MI} \ln PM1 + \alpha_2^{MI} \ln PM2 + \alpha_3^{MI} \ln PM3 \\ + \alpha_4^{MI} \ln PM4 + \alpha_5^{MI} \ln PM5 \qquad (22) \\ + \frac{1}{2} [\beta_{11}^{EI} (\ln PM1)^2 + \beta_{12}^{EI} \ln PM1 \ln PM2 + \ldots], \\ (I = 1, 2 \ldots 9),\end{aligned}$$

where *PM*1 is the price of agriculture, nonfuel mining, and construction, *PM*2 the price of manufacturing, excluding petroleum refining, *PM*3 the price of transportation, *PM*4 the price of communications, trade, and services, and *PM*5 the price of competitive imports.

For both energy (*E*) and materials (*M*) submodels we can derive a system of five equations for determining the relative shares of the five commodity groups making up each submodel. Each equation gives the relative share of one of the commodity groups as a function of the prices of all five groups included in the submodel. We can derive the relative demand functions for each commodity group by dividing the relative value share of the group by the ratio of the price of that group to the price of the corresponding aggregate. For example, to derive the demand for coal relative to total energy we divide the relative value share of coal by the ratio of the price of coal to the price of total energy. We can derive the input-output coefficient for coal by multiplying the demand for coal relative to total energy by the demand for energy relative to the output of the corresponding industrial sector.

The value of each aggregate is equal to the sum of the values of the commodity groups that make up that aggregate. For example, the value of energy is equal to the sum of the values of the five types of energy:

$$\begin{aligned} PE*EI = PE1*E1I + PE2*E2I + PE3*E3I \\ + PE4*E4I + PE5*E5I, \qquad (23) \\ (I = 1, 2 \ldots 9), \end{aligned}$$

where $E1I$ is the quantity of coal, $E2I$ the quantity of crude petroleum and natural gas, $E3I$ the quantity of refined petroleum products, $E4I$ the quantity of electricity, and $E5I$ the quantity of gas as a product of gas utilities. As before, the relative shares of the five energy inputs add to unity, so that the parameters of the five relative demand functions for these inputs must satisfy restrictions analogous to the restrictions given above for the parameters of the aggregate (*KLEM*) submodel. Similar restrictions hold for the five relative demand functions for nonenergy inputs.

☐ **Parameter estimation.** For each of the nine industrial sectors included in our interindustry model of the production structure the aggregate (*KLEM*) submodel consists of four equations. We fit the three equations for relative shares of capital (K), labor (L), and energy (E). The relative shares of materials (M) can be determined from these three equations and the accounting identity between the value of output and the value of input. Taking into account the symmetry restrictions on the parameters of the three equations of the aggregate (*KLEM*) submodel, the number of unknown parameters to be estimated is reduced to nine. Taking convexity restrictions into account where appropriate, we further reduce the number of unknown parameters.[17]

For four of the industrial sectors included in our interindustry model the energy (E) submodel consists of five equations for the relative shares of coal, crude petroleum and natural gas, refined petroleum products, electricity, and gas as a product of gas utilities. These four sectors are:

(1) agriculture, nonfuel mining, and construction
(2) manufacturing, excluding petroleum refining
(4) communication, trade, and services
(7) petroleum refining.

For these industrial sectors we fit the four equations for relative shares of coal, crude petroleum and natural gas, refined petroleum products, and electricity. The relative share of gas as a product of gas utilities can be determined from these four equations and the accounting identity between the total value of energy and the sum of the values of the five types of energy.

For the four industrial sectors listed above the energy (E) submodel involves six symmetry restrictions in the four equations we estimate directly and four additional restrictions that are implicit in these equations and the accounting identity for the total value of energy. In the absence of these restrictions and the restrictions implied by the accounting identity, the energy (E) submodel for the four industrial sectors involves unknown parameters. Taking these restrictions into account, we reduce the number of unknown parameters to fourteen.

---

[17] Methods for imposing convexity restrictions have been developed by Lau [28].

For four of the industrial sectors included in our interindustry model the energy (E) submodel consists of four equations for relative shares of four types of energy. For transportation, coal mining, and electric utilities the relative share of crude petroleum and natural gas is zero. For gas utilities the relative share of electricity is zero. For these four sectors the form of the energy (E) submodel is analogous to the form of the aggregate ($KLEM$) submodel. We fit three equations for the relative shares of three types of energy, excluding the equation for the relative share of gas as a product of gas utilities. The relative share of gas can be determined from the accounting identity for the total value of energy. We fit these three equations subject to symmetry restrictions so that the number of unknown parameters is reduced to nine.

For the crude petroleum and natural gas sector of our interindustry model the energy (E) submodel consists of three equations for relative shares of three types of energy. For this sector the relative shares of coal and gas as an output of gas utilities are equal to zero. We fit two equations for the relative shares of crude petroleum and natural gas and refined petroleum products. The relative share of electricity can be determined from the accounting identity for the total value of energy. Fitting these equations subject to symmetry restrictions, we reduce the number of unknown parameters to five.

For the four nonenergy sectors included in our interindustry model the materials (M) submodel consists of five equations for the relative shares of agriculture, manufacturing, transportation, communications, and competitive imports. For the five energy sectors the materials submodel consists of four equations for the relative shares of agriculture, manufacturing, transportation, and communications. The form of the materials (M) submodel for the four nonenergy sectors is analogous to the form of the energy (E) submodel with five equations for the relative shares of five types of energy. For these sectors we fit four equations for the relative shares of nonenergy inputs, excluding competitive imports. The relative shares of competitive imports can be determined from these four equations and the accounting identity between the total value of materials and the sum of the values of the five types of nonenergy inputs. Each of these submodels involves fourteen unknown parameters.

For the five nonenergy sectors we fit three equations for the relative shares of nonenergy inputs, excluding inputs of communications, trade, and services. The materials (M) submodels for these sectors are analogous to the aggregate ($KLEM$) submodel and involve nine unknown parameters.

For each of the nine industrial sectors included in our interindustry model of production all three submodels—aggregate

(*KLEM*), energy (*E*), and materials (*M*)—have been fitted to annual data on interindustry transactions, capital and labor services, and competitive imports for the period 1947 to 1971.[18] Our method to estimation is the minimum distance estimator for nonlinear simultaneous equations, treating the prices of competitive imports as exogenous variables[19] For each of these submodels the system of equations is nonlinear in the variables but linear in the parameters.

In Tables 1, 2, and 3 we present estimates of the parameters of the translog price possibility frontier for each of the three submodels of our econometric model of producer behavior for all nine industrial sectors. For convenience of the reader we list again the nine sectors:

(1) agriculture, nonfuel mining, and construction
(2) manufacturing, excluding petroleum refining
(3) transportation
(4) communications, trade, and services
(5) coal mining
(6) crude petroleum and natural gas
(7) petroleum refining
(8) electric utilities
(9) gas utilities.

Table 1 contains estimates of the parameters of the translog

TABLE 1

ESTIMATES OF THE PARAMETERS OF THE TRANSLOG PRICE POSSIBILITY FRONTIER FOR THE AGGREGATE (KLEM) SUBMODEL FOR NINE INDUSTRIAL SECTORS OF THE U.S ECONOMY, 1947-1971

	SECTORS								
PARAMETER	1	2	3	4	5	6	7	8	9
α^{I}_{K}	0.1785	0.1149	0.1799	0.2994	0.1277	0.4272	0.1373	0.3458	0.2165
α^{I}_{L}	0.2354	0.2940	0.4096	0.4171	0.4139	0.0987	0.0978	0.1925	0.1085
α^{I}_{E}	0.0244	0.0202	0.0380	0.0182	0.1857	0.1101	0.4553	0.2120	0.5547
α^{I}_{M}	0.5616	0.5708	0.3726	0.2653	0.2727	0.3640	0.3096	0.2496	0.1204
β^{I}_{KK}	0.0851	0.0590	0.1018	0.0595	0.0280	0.2447	0.0849	0.1330	0.0749
β^{I}_{KL}	-0.0366	0.0030	-0.0601	0.0114	-0.0357	-0.0422	-0.0242	0.0288	-0.1181
β^{I}_{KE}	-0.0052	-0.0055	-0.0137	0.0011	0.0099	-0.0470	-0.0772	-0.1682	-0.0941
β^{I}_{KM}	-0.0434	-0.0565	-0.0280	-0.0719	-0.0022	-0.1555	0.0165	0.0064	0.1373
β^{I}_{LL}	0.0287	0.0737	0.0582	0.0848	-0.0751	-0.0131	-0.0174	-0.0968	-0.2773
β^{I}_{LE}	0.0023	0.0054	0.0180	0.0098	0.1145	0.0459	-0.0122	0.0239	0.1318
β^{I}_{LM}	0.0056	-0.0821	0.0199	-0.1059	-0.0037	0.0093	0.0538	0.0441	0.2636
β^{I}_{EE}	0.0072	0.0188	0.0198	0.0020	0.0087	0.0184	0.2282	0.0638	0.1413
β^{I}_{EM}	-0.0044	-0.0187	0.0119	-0.0129	-0.1332	-0.0173	-0.1388	0.0805	-0.1790
β^{I}_{MM}	0.0422	0.1573	-0.0038	0.1907	0.1392	0.1635	0.0685	-0.1311	-0.2219

---

[18] These data were compiled by Jack Faucett Associates [21].

[19] The minimum distance estimator for nonlinear simultaneous equations is discussed by Malinvaud [32], pp. 325–373.

price possibility frontier for the aggregate submodel. The aggregate price possibility frontier is defined on the prices of capital (K), labor (L), energy (E), and materials (M). For each of the nine industrial sectors the parameters of the aggregate submodel are estimated from a system of three equations. The dependent variables in these equations are the relative shares of capital, labor, and energy in the value of total output. Parameters in the equation for the relative share of materials are estimated from the restrictions implied by the accounting identity between the value of output and the value of input. We employ constraints across the equations arising from symmetry restrictions on the price possibility frontier for each sector, reducing the number of parameters to be estimated to nine. We also employ convexity restrictions where appropriate to further reduce the number of parameters to be estimated.

Table 2 contains estimates of the parameters of the translog price possibility frontier for the energy (E) submodel. The energy price possibility frontier gives the price of energy for each sector as a function of the prices of five types of energy inputs. The five types of energy are:

(1) coal
(2) crude petroleum and natural gas
(3) refined petroleum products
(4) electricity
(5) gas as a product of gas utilities.

For each of the nine industrial sectors the parameters of the energy submodel are estimated from a system containing as many as four equations. The dependent variables in these equations are the relative shares of each type of energy in the total value of energy. We employ the restrictions implied by the accounting identity between the total value of energy and the sum of the values of all types of energy, the symmetry restrictions, and, where appropriate, convexity restrictions in reducing the number of parameters to be estimated.

Finally, Table 3 contains estimates of the parameters of the translog price possibility frontier for the materials (M) submodel. For the four nonenergy sectors the materials price possibility frontier is defined on the prices of the five types of nonenergy inputs. These are:

(1) agriculture, nonfuel mining, and construction
(2) manufacturing, excluding petroleum refining
(3) transportation
(4) communications, trade, and services
(5) competitive imports.

For the five energy sectors the materials price possibility frontier is defined on the prices of four types of nonenergy inputs, excluding competitive imports. The dependent variables are relative shares of each type of nonenergy input. We employ restrictions on the parameters that are analogous to the restrictions used for the energy submodel.

TABLE 2

ESTIMATES OF THE PARAMETERS OF THE TRANSLOG PRICE POSSIBLITY FRONTIER FOR THE ENERGY (E) SUBMODEL FOR NINE INDUSTRIAL SECTORS OF THE U S ECONOMY, 1947-1971

	SECTORS								
PARAMETER	1	2	3	4	5	6	7	8	9
α_1^{EI}	0 0053	0 2040	0 0799	0 1142	0 8510	0 0738	0 0448	0 3165	0 0909
α_2^{EI}	0 0021	0 0002	0 0000	0 0111	0 0000	0 1464	0 2131	0 0000	0 1408
α_3^{EI}	0 8389	0 3384	0.8107	0 3520	0 3997	0 0894	0 2289	0 1173	0 0940
α_4^{EI}	0 1212	0 2858	0 0406	0 4136	0 1062	0 6904	0 5132	0 3829	0 6743
α_5^{EI}	0 0325	0 1716	0.0688	0 1091	0 0029	0 0000	0 0000	0 1829	0 0000
β_{11}^{EI}	0 0052	0 1624	0 0735	0 1011	-0 0118	-0 0557	-0 0165	0 0762	-0 0752
β_{12}^{EI}	0.0000	0 0000	0 0000	-0 0013	0 0000	0 1098	0.0483	0 0000	0 1850
β_{13}^{EI}	-0 0068	-0 0690	-0.0648	-0 0402	0 0220	-0 0255	0 0005	0.0833	-0.0192
β_{14}^{EI}	0 0011	-0 0583	-0.0032	-0 0472	-0 0100	-0 0286	-0 0323	-0 0578	-0 0906
β_{15}^{EI}	0 0005	-0.0350	-0 0055	-0.0125	-0 0002	0 0000	0 0000	-0 1017	0 0000
β_{22}^{EI}	0.0010	-0 0029	0.0000	0 0110	0 0000	0 0077	0 1113	0 0000	-0 1416
β_{23}^{EI}	0 0007	-0 0001	0.0000	-0 0039	0 0000	0 0053	-0 0592	0 0000	0 0455
β_{24}^{EI}	-0 0007	0 0073	0 0000	-0 0046	0 0000	-0 1228	-0 1003	0.0000	-0 0889
β_{25}^{EI}	-0 0010	-0 0043	0 0000	-0.0012	0 0000	0 0000	0.0000	0 0000	0 0000
β_{33}^{EI}	-0 0252	0 2239	0 1534	0 2281	-0 0287	-0 0418	0 0680	0 0001	-0 0553
β_{34}^{EI}	0 0128	-0 0967	-0.0329	-0 1456	0 0064	0 0620	-0 0092	-0.0966	0 0289
β_{35}^{EI}	0 1854	-0 0581	-0 0557	-0 0384	0 0003	0 0000	0 0000	0.0162	0 0000
β_{44}^{EI}	-0 0410	0 1868	0 0389	0 2425	0.0055	0 0894	0 1418	0 2077	0 1505
β_{45}^{EI}	0 0278	-0 0390	-0 0028	-0 0451	-0 0019	0 0000	0 0000	-0 0502	0 0000
β_{55}^{EI}	-0 0458	0 1364	0 0640	0 0972	0 0018	0 0000	0.0000	0 1357	0 0000

4. Growth model

■ The second component of our framework for energy policy analysis is a model of long-term U. S. economic growth. A model of U. S. economic growth is required to provide totals for consumption, investment, government and export final demand in our econometric model of interindustry transactions. For this purpose conventional econometric models could be used as an alternative to our macro-econometric growth model. However, a growth model is also required to provide the relative prices of capital and labor services, which enter as primary inputs into the nine sectors of our interindustry model. Prices of primary inputs must be generated within a framework that also incorporates primary input quantities and the prices and quantities of final products. For this purpose a new approach to macro-econometric model-building is required.

Our approach to the explanation of economic growth is closely related to the neoclassical theory of economic growth.[20] The build-

---

[20] See, for example, the discussion of the neoclassical two-sector growth

TABLE 3

ESTIMATES OF THE PARAMETERS OF THE TRANSLOG PRICE POSSIBILITY FRONTIER FOR THE MATERIALS (M) SUBMODEL FOR NINE INDUSTRIAL SECTORS OF THE U S. ECONOMY, 1947-1971

	SECTORS								
PARAMETER	1	2	3	4	5	6	7	8	9
α_1^{MI}	0.2578	0.1348	0.1221	0.0819	0.0193	0.0779	0.0562	0.1134	0.1759
α_2^{MI}	0.3777	0.5933	0.1373	0.2548	0.4270	0.0869	0.1656	0.1046	0.1546
α_3^{MI}	0.0653	0.0472	0.1932	0.0532	0.0675	0.0501	0.1736	0.1200	0.1453
α_4^{MI}	0.2674	0.1643	0.4382	0.5774	0.4839	0.5517	0.4659	0.6620	0.4426
α_5^{MI}	0.0318	0.0603	0.1091	0.0327	0.0023	0.2334	0.1388	0.0000	0.0815
β_{11}^{MI}	0.0799	0.0376	0.1072	-0.0454	0.0190	0.0718	0.0530	0.0170	-0.0407
β_{12}^{MI}	-0.1012	-0.0200	-0.0168	0.0848	-0.0083	-0.0068	-0.0093	0.0755	0.0845
β_{13}^{MI}	0.0629	0.0043	-0.2360	-0.0094	-0.0013	-0.0039	-0.0097	-0.0292	-0.0703
β_{14}^{MI}	-0.0672	-0.0571	-0.0535	-0.0806	-0.0094	-0.0430	-0.0262	-0.0633	-0.0899
β_{15}^{MI}	0.0256	0.0352	-0.0133	0.0506	-0.0000	-0.0182	-0.0078	0.0000	0.1164
β_{22}^{MI}	0.2349	0.1958	0.1185	0.0973	0.2447	0.0794	0.1382	0.0023	-0.0070
β_{23}^{MI}	-0.0219	-0.0361	-0.0265	-0.0091	-0.0288	-0.0044	-0.0288	0.0037	-0.0123
β_{24}^{MI}	-0.1009	-0.0710	-0.0602	-0.1179	-0.2066	-0.0480	-0.0772	-0.0815	-0.0895
β_{25}^{MI}	-0.0109	-0.0687	-0.0150	-0.0550	-0.0010	-0.0203	-0.0230	0.0000	0.0243
β_{33}^{MI}	0.0039	0.0435	0.1559	0.0502	0.0629	0.0476	0.1434	0.1027	0.1095
β_{34}^{MI}	-0.0187	-0.0030	-0.0847	-0.0321	-0.0327	-0.0276	-0.0809	-0.0773	-0.0739
β_{35}^{MI}	-0.0263	-0.0087	-0.0211	0.0005	-0.0002	-0.0117	-0.0241	0.0000	0.0471
β_{44}^{MI}	0.1959	0.1218	0.2462	0.2348	0.2497	0.2473	0.2488	0.2221	0.2346
β_{45}^{MI}	-0.0090	0.0093	-0.0478	-0.0042	-0.0011	-0.1287	-0.0646	0.0000	0.0187
β_{55}^{MI}	0.0206	0.0329	0.0972	0.0081	0.0023	0.1789	0.1195	0.0000	-0.2065

ing blocks of our model are submodels of household and production sectors and submodels of foreign and government sectors. Economic growth results from the link between current capital formation and future productive capacity. In our model this link is provided by a macro-econometric production function, relating the output of consumption and investment goods to the input of capital and labor services. Preferences between present and future consumption, which determine the allocation of income between saving and consumption, complete our model of economic growth.

Our macro-econometric growth model of the U. S. economy provides for the simultaneous determination of the values of products and factors of production in both current and constant prices. The model links demand for capital formation by savers to the

---

model by Burmeister and Dobell [7] and the references given there. A more detailed discussion of our model is presented in Hudson and Jorgenson [25]; see also, Jorgenson, Berndt, Christensen, and Hudson [26], Chapter 2.

supply of investment goods by producers. Similarly, the model links demand for consumption goods and supply of labor services by households to supply of consumption goods and demand for labor by producers. Finally, given the supply of capital stock, the demand for capital services determines the overall rate of return to capital.

The theory of U. S. economic growth that underlies our macroeconometric growth model is a theory of the behavior of the private sector of the U. S. economy. The behavior of the foreign and government sectors is taken to be exogenous. Demographic trends —the growth of population, labor force, and unemployment—are also exogenous to the model. The main determinant of growth in productivity is capital formation. Growth in productivity over and above growth due to capital formation is exogenous to the model. We have projected demographic trends and trends in productivity growth on the basis of post-war experience in the United States.[21]

☐ **Variables.** Our econometric growth model is summarized in the following series of tables. In Table 4 we present our notation for the variables that appear in the model. The first group of variables converts aggregates from one basis of classification to another. For example, the index of total factor productivity A converts input to output. The index of AW converts investment weights for capital formation to the weights appropriate for the measurement of wealth. All of the aggregation variables are taken to be exogenous.

The second group of variables appearing in Table 4 comprises the quantities of products and factors of production, broken down by sector of origin and destination. Variables beginning with C are quantities of consumption goods. Similarly, variables beginning with I are quantities of investment goods. Variables beginning with L are quantities of labor services, while variables beginning with K are quantities of capital services. The third group of variables includes prices corresponding to the quantities of products and factors of production. Each price begins with P and continues with the corresponding quantity. For example, the variable C is personal consumption expenditures and the variable PC is the price of personal consumption expenditures.

The fourth group of variables consists of financial variables: rates of depreciation and replacement, the nominal rate of return, gross private national saving, and private national wealth. Finally, the fifth group of variables comprises tax and transfer variables. The variable EL represents government transfer payments to persons other than social insurance funds, an expenditure category. The variables beginning with T are tax rates. Each of the products and factors included in the model—consumption goods, investment goods, capital services, and labor services—is associated with an effective tax rate. The variable TP is the effective tax rate for capital stock.

☐ **Equations.** In Table 5 we present the equations for our macroeconometric growth model. The model includes five behavioral equations, describing the behavior of household and business sec-

---

[21] Detailed projections are presented by Hudson and Jorgenson [25].

TABLE 4

MACRO-ECONOMIC GROWTH MODEL· NOTATION

1 AGGREGATION VARIABLES			
A	TOTAL FACTOR PRODUCTIVITY (INPUT TO OUTPUT)	AK	CAPITAL STOCK, LAGGED TO CAPITAL SERVICE
ACI	INVESTMENT TO CHANGE IN BUSINESS INVENTORIES, CONSUMPTION GOODS	APC	IMPLICIT DEFLATOR OF CONSUMPTION GOODS TO IMPLICIT DEFLATOR OF CHANGE IN BUSINESS INVENTORIES, CONSUMPTION GOODS
AI	INVESTMENT TO CAPITAL STOCK	AW	INVESTMENT TO WEALTH
AL	INVESTMENT TO CAPITAL STOCK, LAGGED		
2 QUANTITIES			
C	PERSONAL CONSUMPTION EXPENDITURES, INCLUDING SERVICES OF CONSUMERS' DURABLES	IG	GOVERNMENT PURCHASES OF INVESTMENT GOODS
CE	SUPPLY OF CONSUMPTION GOODS BY GOVERNMENT ENTERPRISES	IR	NET EXPORTS OF INVESTMENT GOODS
CG	GOVERNMENT PURCHASES OF CONSUMPTION GOODS	IS	SUPPLY OF INVESTMENT GOODS BY PRIVATE ENTERPRISES.
CI	CHANGE IN BUSINESS INVENTORIES OF CONSUMPTION GOODS	L	SUPPLY OF LABOR SERVICES
CR	NET EXPORTS OF CONSUMPTION GOODS, LESS INCOME ORIGINATING, REST OF THE WORLD.	LD	PRIVATE PURCHASES OF LABOR SERVICES
CS	SUPPLY OF CONSUMPTION GOODS BY PRIVATE ENTERPRISES	LGE	GOVERNMENT ENTERPRISES PURCHASES OF LABOR SERVICES
G	NET CLAIMS ON GOVERNMENT	LGG	GENERAL GOVERNMENT PURCHASES OF LABOR SERVICES
R	NET CLAIMS ON REST OF THE WORLD	LH	TIME AVAILABLE
I	GROSS PRIVATE DOMESTIC INVESTMENT, INCLUDING PURCHASES OF CONSUMERS' DURABLES	LJ	LEISURE TIME.
		LR	NET EXPORTS OF LABOR SERVICES
		LU	UNEMPLOYMENT
		K	CAPITAL STOCK
		KD	CAPITAL SERVICES
3 PRICES			
PC	IMPLICIT DEFLATOR, PERSONAL CONSUMPTION EXPENDITURES, INCLUDING SERVICES OF CONSUMERS' DURABLES	PIG	IMPLICIT DEFLATOR, GOVERNMENT PURCHASES OF INVESTMENT GOODS
PCE	IMPLICIT DEFLATOR, SUPPLY OF CONSUMPTION GOODS BY GOVERNMENT ENTERPRISES	PIR	IMPLICIT DEFLATOR, NET EXPORTS OF INVESTMENT GOODS
PCG	IMPLICIT DEFLATOR, GOVERNMENT PURCHASES OF CONSUMPTION GOODS	PIS	IMPLICIT DEFLATOR, SUPPLY OF INVESTMENT GOODS BY PRIVATE ENTERPRISES
PCI	IMPLICIT DEFLATOR, CHANGE IN BUSINESS INVENTORIES OF CONSUMPTION GOODS	PL	IMPLICIT DEFLATOR, SUPPLY OF LABOR SERVICES
PCR	IMPLICIT DEFLATOR, NET EXPORTS OF CONSUMPTION GOODS, LESS INCOME ORIGINATING, REST OF THE WORLD	PLD	IMPLICIT DEFLATOR, PRIVATE PURCHASES OF LABOR SERVICES
PCS	IMPLICIT DEFLATOR, SUPPLY OF CONSUMPTION GOODS BY PRIVATE ENTERPRISES	PLGE	IMPLICIT DEFLATOR, GOVERNMENT ENTERPRISES PURCHASES OF LABOR PURCHASES
PG	IMPLICIT DEFLATOR, NET CLAIMS ON GOVERNMENT	PLGG	IMPLICIT DEFLATOR, GENERAL GOVERNMENT PURCHASES OF LABOR SERVICES
PR	IMPLICIT DEFLATOR, NET CLAIMS ON REST OF THE WORLD	PLR	IMPLICIT DEFLATOR, NET EXPORTS OF LABOR SERVICES
PI	IMPLICIT DEFLATOR, GROSS PRIVATE DOMESTIC INVESTMENT, INCLUDING PURCHASES OF CONSUMERS' DURABLES	PKD	IMPLICIT DEFLATOR, CAPITAL SERVICES
4 FINANCIAL VARIABLES			
D	RATE OF DEPRECIATION, PRIVATE DOMESTIC TANGIBLE ASSETS	N	NOMINAL RATE OF RETURN, PRIVATE DOMESTIC TANGIBLE ASSETS
M	RATE OF REPLACEMENT, PRIVATE DOMESTIC TANGIBLE ASSETS	S	GROSS PRIVATE NATIONAL SAVING.
		W	PRIVATE NATIONAL WEALTH
5 TAX AND TRANSFER VARIABLES			
EL	GOVERNMENT TRANSFER PAYMENTS TO PERSONS, OTHER THAN FROM SOCIAL INSURANCE FUNDS	TK	EFFECTIVE TAX RATE, CAPITAL SERVICES
TC	EFFECTIVE TAX RATE, CONSUMPTION GOODS	TL	EFFECTIVE TAX RATE, LABOR SERVICES.
TI	EFFECTIVE TAX RATE, INVESTMENT GOODS	TP	EFFECTIVE TAX RATE, CAPITAL STOCK

tors.[22] The demand for labor is determined by the total level of production, the amount of capital services available, and the relative

---

[22] Our model of the household sector was originated by Christensen and Jorgenson [12]. Our model of the business sector was originated by Christensen, Jorgenson, and Lau [18].

TABLE 5

MACRO-ECONOMIC GROWTH MODEL: EQUATIONS

1 BEHAVIORAL EQUATIONS

INVESTMENT SUPPLY:

$$\frac{PIS*IS}{PKD*KD} = 1.1717 - 0.5006 * (\text{LOG } CS - \text{LOG } IS)$$

LABOR DEMAND,

$$\frac{PLD*LD}{PKD*KD} = 1.5655$$

PRODUCTION POSSIBILITY FRONTIER:

$$0 = -\text{LOG } KD - 1.5655 * \text{LOG } LD + 1.3938 * \text{LOG } CS - 2.5655 * \text{LOG } A$$
$$+ 1.1717 * \text{LOG } IS + 0.2503 * (\text{LOG } CS - \text{LOG } IS) ** 2$$

CONSUMPTION DEMAND

$$PC * C = 0.0034 * W(-1) + 0.1469 * (PL * LH + EL)$$

LEISURE DEMAND.

$$PL * LJ = 0.0196 * W(-1) + 0.8403 * (PL * LH + EL)$$

2 ACCOUNTING IDENTITIES

CAPITAL STOCK AND INVESTMENT.

$$K = AI * I + (1 - M) * K(-1)$$

CAPITAL SERVICE AND CAPITAL STOCK

$$KD = AK * K(-1).$$

VALUE OF OUTPUT AND INPUT:

$$PIS * IS + PCS * CS = PKD * KD + PLD * LD.$$

VALUE OF CONSUMPTION GOODS

$$(1 + TC) * PCS * CS + PCE * CE$$
$$= PC * C + PCG * CG + PCI * CI + PCR * CR$$

VALUE OF INVESTMENT GOODS.

$$(1 + TI) * PIS * IS + PCI * CI = PI * I + PIG * IG + PIR * IR$$

VALUE OF CAPITAL SERVICES.

$$(1 - TK) * (PKD * KD - TP * PI[-1] * AW[-1] * K[-1])$$
$$= N * PI(-1) * AW(-1) * K(-1) + D * PI * AL * K(-1)$$
$$+ PI(-1) * AW(-1) * K(-1) - PI * AL * K(-1)$$

VALUE OF LABOR SERVICES

$$(1 - TL) * (PLD * LD + PLGE * LGE + PLGG * LGG + PLR * LR) = PL * L$$

SAVING:

$$S = PI * I + PG * (G - G[-1]) + PR * (R - R[-1])$$

WEALTH

$$W = PI * AW * K + PG * G + PR * R$$

3 BALANCE EQUATIONS

CONSUMPTION:

$$CS + CE = C + CG + CI + CR$$

INVESTMENT.

$$IS + CI = I + IG + IR$$

TIME

$$LH = L + LJ.$$

LABOR

$$L = LD + LGE + LGG + LR + LU$$

4 AGGREGATION EQUATIONS

IMPLICIT DEFLATOR, CHANGE IN BUSINESS INVENTORIES, CONSUMPTION GOODS

$$PCI = PC * APC$$

CHANGE IN BUSINESS INVENTORIES, CONSUMPTION GOODS:

$$PCI * CI = PI * I * ACI$$

prices of capital and labor services. Output of investment goods is determined by the price of investment goods, the prices of capital services and the available supply of capital services, and the amount of productive capacity being devoted to the output of consumer goods and services. The production that takes place in the U. S. private sector, whether of consumption or of investment goods, is limited by the total productive capacity, which in turn depends on available supplies of capital and labor services as well as on the level of technology.

The level of household expenditure on consumer goods and services is determined by the wealth and resources held by the household sector, including the time endowment of the sector. The desired amount of work input provided by the household sector is determined by the total amount of time available, the wage rate, and the extent of other resources available to the household sector in the form of wealth and transfer payments.

The behavioral equations of our macro-econometric growth model have been estimated from historical data for the United States for the period 1929 to 1971.[23] In addition to the five behavioral equations the model includes accounting identities for capital stock, investment and capital services, for the value of input and output, for saving and wealth, and for the value of consumption goods, investment goods, capital services, and labor services. These accounting identities incorporate the budget constraints for household and business sectors and the flow of each product and factor of production in current prices.

The model is completed by balance between demand and supply of products and factors of production in constant prices and by aggregation equations that determine inventory accumulation of consumption goods. Although gross private domestic investment is determined in the model, the allocation of investment between fixed investment and inventory accumulation is not determined in the model. An allocation between inventory accumulation in the form of consumption goods and other components of gross private domestic investment is required for the balance between demand and supply of consumption and investment goods. In Figure 3 we outline the working of our macro-econometric growth model by means of a diagram.

5. Energy projections

■ Our next objective is to provide a reference point for the analysis of energy policy by establishing detailed projections of energy demand and supply, energy price and cost, and energy imports and exports. Our projections cover the years 1975, 1980, 1985, and 2000 and are based on the assumption that there are no major changes in energy policy, either by U. S. or foreign governments, over the forecast period. The projections include the entire matrix of interindustry transactions in current dollar and in constant dollar flow. We translate this information into physical terms by converting the constant dollar energy transactions into British thermal units (Btu) for each fuel.[24] We also convert the price indices into dollars per physical unit. This transformation permits the expression

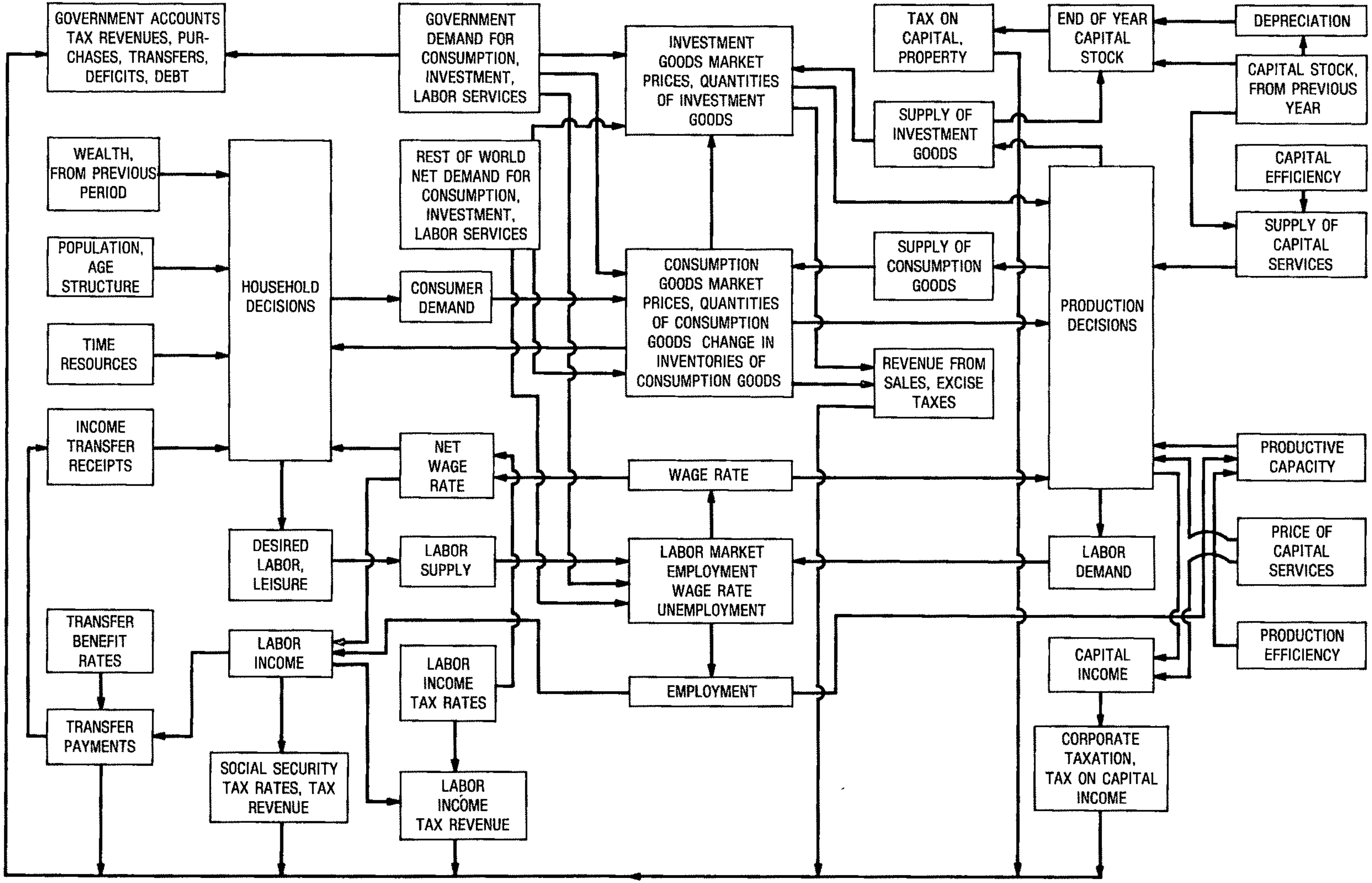
---

[23] The data are presented in a series of articles by Christensen and Jorgenson [13, 14, 15, 16].

[24] The conversion process is discussed in the following section.

FIGURE 3

MACRO-ECONOMETRIC GROWTH MODEL DIAGRAMMATIC REPRESENTATION


GOVERNMENT ACCOUNTS TAX REVENUES, PURCHASES, TRANSFERS, DEFICITS, DEBT
WEALTH, FROM PREVIOUS PERIOD
POPULATION, AGE STRUCTURE
TIME RESOURCES
INCOME TRANSFER RECEIPTS
TRANSFER BENEFIT RATES
TRANSFER PAYMENTS
HOUSEHOLD DECISIONS
DESIRED LABOR, LEISURE
LABOR INCOME
SOCIAL SECURITY TAX RATES, TAX REVENUE
GOVERNMENT DEMAND FOR CONSUMPTION, INVESTMENT, LABOR SERVICES
REST OF WORLD NET DEMAND FOR CONSUMPTION, INVESTMENT, LABOR SERVICES
CONSUMER DEMAND
NET WAGE RATE
LABOR SUPPLY
LABOR INCOME TAX RATES
LABOR INCOME TAX REVENUE
INVESTMENT GOODS MARKET PRICES, QUANTITIES OF INVESTMENT GOODS
CONSUMPTION GOODS MARKET PRICES, QUANTITIES OF CONSUMPTION GOODS CHANGE IN INVENTORIES OF CONSUMPTION GOODS
WAGE RATE
LABOR MARKET EMPLOYMENT WAGE RATE UNEMPLOYMENT
EMPLOYMENT
TAX ON CAPITAL, PROPERTY
SUPPLY OF INVESTMENT GOODS
SUPPLY OF CONSUMPTION GOODS
REVENUE FROM SALES, EXCISE TAXES
END OF YEAR CAPITAL STOCK
PRODUCTION DECISIONS
LABOR DEMAND
CAPITAL INCOME
CORPORATE TAXATION, TAX ON CAPITAL INCOME
DEPRECIATION
CAPITAL STOCK, FROM PREVIOUS YEAR
CAPITAL EFFICIENCY
SUPPLY OF CAPITAL SERVICES
PRODUCTIVE CAPACITY
PRICE OF CAPITAL SERVICES
PRODUCTION EFFICIENCY

of our detailed interindustry projections in the energy balance framework that is conventional in energy analysis.[25]

A summary of the composition and growth of the interindustry transactions for the U. S. economy is given in Table 6. This table presents information on the gross output of each producing sector, together with information on the disposition side of GNP. The rate of growth of real GNP is expected to slow somewhat from recent levels, in large part because of the expected decline in the rate of increase of the labor force, but only to around 3.85 percent a year. Inflation also is expected to slow from rates experienced in the recent past but, at 3.76 percent a year, to remain above typical historical rates.

TABLE 6

U.S. INDUSTRY OUTPUT AND PRICES, 1975-2000 (AVERAGE ANNUAL GROWTH RATES, PERCENT)

OUTPUT OF SECTOR	REAL DOLLARS	PRICES	CURRENT DOLLARS
1	2.82	4.97	7.93
2	3.92	3.49	7.55
3	4.30	2.33	6.73
4	3.51	4.15	7.81
5	2.73	6.45	9.36
6	1.64	4.46	6.17
7	2.15	5.80	8.06
8	5.25	3.54	8.07
9	1.05	6.68	8.98
CONSUMPTION	3.82	3.64	7.80
INVESTMENT	3.80	3.93	7.88
GOVERNMENT	3.78	4.11	8.05
NET EXPORTS	6.29		
GNP	3.85	3.76	7.75

The composition of GNP is expected to change gradually. Net exports absorb an increasing fraction of GNP as the terms of trade, particularly relating to raw materials, continue to move against the United States. Real government purchases fall in relation to total output, although the rapid rate of increase in the price of government purchases, primarily of its purchases of labor and services, offsets this and results in a small increase in the current dollar share of government purchases of GNP. Personal consumption in real terms increases in line with GNP but, because of the slower than average increase in consumption prices, the current dollar share of consumption in GNP falls. Private investment in real terms increases more slowly than GNP but, in current dollars, it increases more rapidly.

The composition of production changes more noticeably. Agriculture, nonfuel mining, and construction output increases relatively slowly, as the output is income inelastic, while the expected productivity advance in manufacturing and transport permits a comparatively rapid increase in output from these sectors with a less than average increase in prices. Communications, trade, and services output continues to increase but less rapidly than real GNP and, due to the slow productivity advance in this sector and its relative intensity in an increasing cost input, labor, its prices increase comparatively rapidly.

---

[25] See note 5, above.

☐ **Energy balance.** Our projection of total U. S. energy utilization for the period 1975 to 2000 is summarized in Table 7. Total energy input is forecast to increase from 77.099 quadrillion Btu in 1975 to 161.241 quadrillion Btu in 2000, an average annual growth rate of 3.0 percent. The rate of increase of this total varies over time. In the past the rate of increase of total input has varied from 3.7 percent a year over the 1960 to 1965 period, to 4.8 percent over 1965 to 1970, to 3.9 percent for 1970 to 1973. Our forecast 1975 total corresponds to a 2.7 percent increase for the 1970 to 1975 period.

TABLE 7

U.S ENERGY INPUT, 1970-2000

	1970	1975	1980	1985	2000
1 U S TOTAL ENERGY INPUT (QUADRILLION BTU)					
COAL	12.922	13 260	14 304	16 538	26 891
PETROLEUM	29 614	33 696	39 176	44 547	58 148
NATURAL GAS (EXC.SYN.GAS)	22 029	24 727	26 888	28 145	28 681
HYDRO, NUCLEAR, OTHER	2 879	5 416	10 115	16 096	47 521
TOTAL	67.444	77 099	90 484	105.326	161.241
2 TOTAL ENERGY INPUT, GROWTH RATES (AVERAGE PERCENT PER ANNUM)					
COAL		0 52	1 52	2 94	3 29
PETROLEUM		2 62	3 06	2 60	1 79
NATURAL GAS		2.34	1 69	0 92	0 13
HYDRO, NUCLEAR, OTHER		13 47	13 31	9 74	7 48
TOTAL		2 71	3 25	3 08	2 88
3 TOTAL ENERGY INPUT, COMPOSITION (PERCENT OF TOTAL)					
COAL	19 16	17 20	15 81	15 70	16 68
PETROLEUM	43 91	43 70	43 30	42 29	36 06
NATURAL GAS	32 66	32 07	29 72	26 72	17 79
HYDRO, NUCLEAR, OTHER	4 27	7 02	11 18	15 28	29 47
TOTAL	100 00	100 00	100 00	100 00	100 00

The continued increase in fuel prices and the continued presence of some general interest in energy conservation, together with the changing structure of the economy, prevent the rate of increase in energy use from returning to historical trends. These forces lead to several avenues of energy conservation. Wasteful energy use is reduced, existing capital is gradually replaced with more energy efficient capital, and output shifts towards less energy-intensive forms, leading to a steady decline in the projected rate of increase of total energy input. Thus, projected U. S. gross energy input increases at 3.2 percent over the 1975 to 1980 period, at 3.3 percent over 1980 to 1985, and at 2.9 percent over 1985 to 2000.

The composition of total energy input is also expected to change markedly. Coal is projected to decline slightly in relative importance until 1985 but then to increase until 2000 as new demand for coal for synthetic gas production is superimposed on the continuing demand for coal for electricity generation. The net result is that the share of coal in total input is almost the same in 2000 as in 1975. Petroleum shows a continuing decline in relative importance. Use of petroleum continues to increase at around 2.8 percent a year until 1985 but this increase gradually slows as the introduction of more energy efficient capital and change in economic patterns have their principal effects in reducing demand for petroleum. For example, use of more efficient automobiles, more use of public transport, better building insulation, use of heat pumps in heating and cooling all have a major effect in slowing the increase in petroleum use. Thus, by 2000, petroleum is projected to form 36 percent of total energy input, compared to 44 percent in the 1970s.

Natural gas is predicted to decline dramatically in relative importance. This is due primarily to expected supply limitations which prevent its use from keeping pace with the other fuels. The share of natural gas in total energy falls from 32 percent in 1975 to 18 percent in 2000 (although, in terms of consumption, synthetic gas supplements the availability of gaseous fuels). Finally, the hydro, nuclear, and other share in total input is expected to increase dramatically. This is due to rapidly increasing use of electricity and to the steadily increasing importance of nuclear generation within the electricity sector. The 9-percent average annual increase of hydro and nuclear input over the 1975 to 2000 period results in its share in total input's increasing from 7 percent to 29 percent.

Consumption of each type of energy is shown in Table 8. Coal consumption increases at an average rate of 2.7 percent over the 1975 to 2000 period, petroleum consumption at a rate of 2.2 percent, electricity at 5.3 percent and gas (natural plus synthetic) at 1.1 percent. These growth rates vary within the period, typically increasing over the 1975 to 1980 period, compared to rates for

TABLE 8

U S ENERGY CONSUMPTION, 1970-2000*

	1970	1975	1980	1985	2000
1 U S ENERGY CONSUMPTION (QUADRILLION BTU)					
COAL	12 922	13 260	14 304	16 538	26.891
PETROLEUM	29.614	33 696	39 176	44 547	58 148
ELECTRICITY	5 218	6 646	8 854	11 499	23 873
GAS	22.029	24 727	26 888	28 835	32 321
2 ENERGY CONSUMPTION, GROWTH RATES (AVERAGE PERCENT PER ANNUM)					
COAL		0 52	1 52	2.94	3 29
PETROLEUM		2 62	3.06	2 60	1 79
ELECTRICITY		4 96	5 90	5 37	4 99
GAS		2 34	1 69	1 41	0 76
3 ENERGY CONSUMPTION, PHYSICAL UNITS					
COAL (MILLION SHORT TONS)	525 00	573 00	618 00	715 00	1163 00
PETROLEUM (MILLION BARRELS A DAY)	14 70	16 72	19 44	22.11	28 86
ELECTRICITY (BILLION KILOWATT HOURS)	1530 00	1938 00	2590 00	3363 00	6981 00
GAS (TRILLION CUBIC FEET)	21 37	23 97	25 90	27 76	31 10

*COAL CONSUMPTION INCLUDES COAL USED IN ELECTRICITY GENERATION AND IN THE PRODUCTION OF SYNTHETIC GAS PETROLEUM CONSUMPTION INCLUDES PETROLEUM INPUTS INTO ELECTRICITY GENERATION AND SYNTHETIC GAS GAS CONSUMPTIONS INCLUDES BOTH NATURAL AND SYNTHETIC GAS

1970 to 1975, which are reduced as a result of the 1973 to 1974 energy crisis. After 1980 the growth rates steadily decline over the remainder of the century. The dominant trend in these aggregate consumption figures is the continued rapid increase in electricity consumption.

The rapid increase in electricity use has several implications. First, since electricity is a secondary form of energy suffering large energy losses in the conversion from primary fuels, it is very expensive in terms of energy input and its rapid growth produces a rapid growth in the use of primary fuels. In other words, the cost of the electricity growth in energy terms is the absorption of ever increasing proportions of total energy input in electricity conversion loss. For example, in 1975 electricity conversion losses are projected to amount to 13.9 quadrillion Btu, or 18.1 percent of total energy input, but in 2000 these losses are projected to be about 50.0 quadrillion Btu, which is 31.0 percent of total energy.

The second implication of electricity use is derived from the fact that any fuel can be used as the primary input to electricity generation. Uranium and hydro resources, at present, have no alternative use, although uranium enrichment for present generation nuclear reactors is a heavy user of electricity. Also, the ability to use any fossil fuel gives electricity generation some degree of comparative advantage over final use of these fuels. Electricity generation can exploit coal, which is in relative abundance in the United States, and residual oil, which, for technical reasons, has only limited value as a fuel in other uses. In short, electricity has a property that partially offsets its large energy requirements, for the cheapest and most abundant fuels can be used in its generation, leaving scarce oil and gas supplies available for direct use.

Projected fuel prices are shown in Table 9. All prices are expected to increase in current dollars, coal by an average of 6.5 percent a year over the 1975 to 2000 period, crude petroleum by

TABLE 9

U.S ENERGY PRICES, 1970-2000

	1970	1975	1980	1985	2000
1 FUEL PRICES, CURRENT DOLLARS					
COAL ($/SHORT TON)	7 59	11 48	16 38	22 84	54 83
CRUDE PETROLEUM ($/BARREL)˙	3 39	7 12	8 64	10 97	21 20
ELECTRICITY (c/KILOWATT HR)	1 61	2 11	2 56	3 01	5 04
GAS ($/THOUSAND CUBIC FT)	0 83	1 05	1 48	1 97	5 03
2 FUEL PRICES, GROWTH RATES (AVERAGE PERCENT PER ANNUM)					
COAL		8 63	7 36	6 88	6 01
CRUDE PETROLEUM		16 00	3 95	4 89	4 49
ELECTRICITY		5 56	3 94	3 29	3 50
GAS		4 81	7 11	5 89	6 45
GNP PRICE DEFLATOR		5 63	4 11	3 70	3 66
3 FUEL PRICES, CONSTANT DOLLARS					
COAL ($1975/SHORT TON)	5 77	11 48	13 39	15 57	21 80
CRUDE PETROLEUM ($1975/BARREL)	4 46	7 12	7 06	7 48	8 42
ELECTRICITY (c1975/KILOWATT HR)	2 12	2 11	2 09	2 06	2 00
GAS($1975/THOUSAND CUBIC FT)	1 09	1 05	1 21	1 34	2 00

4.5 percent, refined petroleum products by 5.8 percent, electricity by 3.5 percent, and gas by 6.5 percent. These rates of increase reflect both supply and demand factors. The rapid increase in coal prices is due largely to the slow productivity advance expected in the coal mining industry; increasing difficulty and expense in securing crude oil both domestically and from foreign sources, combined with demand conditions in petroleum product markets, operate to produce continuing rapid increases in oil prices; electricity prices increase but, due to continued productivity advance in the electric utilities sector, at a much slower rate than other prices; gas prices rise rapidly as an increasing demand faces a relatively inelastic supply (this presupposes a relaxation of price controls on natural gas).

When the rates of increase of fuel prices are compared to general inflation, which averages 3.8 percent over the 1975 to 2000 period, a somewhat different picture emerges. First, coal, petroleum, and gas prices all show a significant increase in real terms, which is in marked contrast to historical experience, e.g. between 1951 and 1971 real coal prices fell 15 percent and real prices for refined petroleum products fell 17 percent. This rising price trend provides a strong incentive for economy in the use of fossil fuels. The second price feature is that electricity prices, in real terms, fall slightly, although the 5-percent fall between 1975 and 2000 is again much less than the 43-percent fall that occurred between 1951 and 1971. Compared to past experience, the slower decline in real electricity prices exerts some pressure to slow the historical rate of increase in electricity use, but the main effect of the real electricity price decline is to continue to promote the substitution of electricity for other fuels.

The fuel supply sections of the energy forecasts are shown in Table 10. Although our interindustry model takes account of production and supply possibilities after the primary extractive stage, it does not include any detailed information on the supply characteristics of the U. S. fuel extracting industries. Indeed, given current knowledge, any supply side predictions, particularly to 2000, are hazardous. The supply figures shown in Table 10 represent, essentially, estimates appended to the model. Coal has no supply problems concerning its existence; the only problems concern the prices at which mining and land reclamation become economic. Our coal supply information represents an extrapolation of past data and is embodied in the coal sector production structure of the model. The U. S. oil output figures are based on an *ad hoc* combination of past price elasticities and current information concerning future oil fields, such as the North Slope and other Alaskan fields. Imports make up the difference between U. S. demand and supply at each oil price. Our estimates for U. S. output of natural gas are based upon estimates made by various government and private agencies, as are our estimates for the output of synthetic gas and of shale oil.

The energy flows forecast for the 1975 to 2000 period are presented in detail in Table 11 and in summary form in Table 12. The interindustry energy flows are shown in trillions of Btu for each of the four fuels—coal, petroleum, electricity, and gas—into

TABLE 10

U.S. ENERGY SUPPLY, 1970-2000

	1970	1975	1980	1985	2000
COAL (MILLION SHORT TONS)					
PRODUCTION	598	658	707	819	1291
EXPORTS	72	85	89	104	129
U S CONSUMPTION	525	573	618	715	1163
PETROLEUM (MILLION BARRELS A DAY)					
U S. CRUDE OUTPUT (INCLUDING GAS LIQUIDS)	11 54	10 61	13 91	17 82	21.41
U S OUTPUT OF SHALE OIL	0 00	0 00	0 00	0 00	4.29
IMPORTS	3 42	6 40	5 78	4 51	3.28
U S CONSUMPTION	14 70	16 72	19 44	22 11	28 86
EXPORTS	0 26	0 29	0.24	0 23	0 12
GAS (BILLION CUBIC FEET)					
U S OUTPUT OF NATURAL GAS	20616	21838	22510	23620	23812
U S OUTPUT OF SYNTHETIC GAS	0	0	0	664	3502
IMPORTS	821	2210	3443	3525	3813
U S CONSUMPTION	21367	23966	25898	27762	31097
EXPORTS	70	81	55	47	30
TOTAL (QUADRILLION BTU)					
U S ENERGY INPUT	67 444	77 099	90 484	105 326	161 241
EXPORTS	2 610	2 604	2.601	2 905	3 256
TOTAL DEMAND	70 054	79 703	93 085	108 231	164 497
IMPORTS	8 235	15 880	15 886	13 273	11 040
SUPPLEMENTAL (SYNTHETIC GAS, SHALE OIL)	0 000	0 000	0 000	0 690	12 281
IMPORTS AS PERCENTAGE OF TOTAL DEMAND	11 76	19.92	17 07	12 26	6 71
SUPPLEMENTAL AS PERCENTAGE OF TOTAL DEMAND	0 00	0 00	0 00	0 64	7 47

each of the nine intermediate and four final use sectors. These Btu flows are obtained by applying exogenous Btu/constant dollar ratios to the fuel entries in the constant dollar transaction matrices. Since we are operating at a greater level of disaggregation of fuel uses than is given in energy data, it was necessary to make some simplifying assumptions to obtain the translation of dollars to Btu's.

Information is available for the Btu/constant dollar ratio for each fuel and this ratio for fuel inputs to electricity generation can also be accurately derived from published data. Thus, there is no problem in deriving the figures for total use of each fuel or for fuel inputs into electricity. To fill in the remaining Btu entries it was assumed that, within each fuel, the same Btu/constant dollar ratio applied to each remaining entry with certain exceptions. The exceptions, which embody what other information is available, are that sales of electricity to households and to agriculture take place at a price 67 percent above that charged other users, that petroleum products to households and to agriculture have, because of the product mix, a price 50 percent and 25 percent, respectively, above the average price charged other users, and that the price of gas to households in 1971 was $1.48/million Btu, the price to manufacturing was $0.56, and that gas prices to other users are equal.

Coal is used primarily in manufacturing, coal mining, electricity generation, and for export. The first two uses are relatively unchanging; the projected growth in coal consumption comes in its use in electricity generation, its use for export, and, in the latter

TABLE 11

COMPOSITION OF U.S. ENERGY USE, 1970-2000, INTERMEDIATE AND FINAL DEMAND BY SECTOR (ALL FIGURES IN TRILLION BTU)

	SUPPLY		TOTAL DEMAND, SUPPLY	DEMAND ENERGY INPUTS TO SECTOR												
	U S OUTPUT	IMPORTS		1	2	3	4	5	6	7	8	9	10	11	12	13
COAL																
1970	14911	2	14913	27	3342	15	440	1223	0	17	7483	6	173	159	36	1991
1975	15208	9	15217	0	3225	0	147	1228	0	0	8522	0	102	4	33	1955
1980	16356	9	16365	0	3040	0	74	853	0	0	10227	0	74	4	33	2061
1985	18929	10	18939	0	2833	0	66	863	0	0	11818	844	75	4	34	2401
2000	29850	16	29866	0	3904	0	90	1316	0	0	15985	5504	32	7	43	2975
PETROLEUM																
1970	22773	7388	30161	3830	4747	2614	3903	32	83	150	2087	148	10523	126	1360	547
1975	20677	13592	34269	4349	5587	2931	4383	34	51	127	2375	165	12115	79	1500	573
1980	27330	12327	39658	4897	6476	3696	5057	59	44	138	2567	197	14261	107	1677	482
1985	35212	9628	45001	5586	7575	4480	5931	94	29	160	2830	318	15555	125	1865	453
2000	51252	7092	58395	8244	11602	6210	9340	165	0	181	3918	625	15366	119	2379	247
ELECTRICITY																
1970	5220	0	5220	101	1388	19	1459	21	44	43	596	0	1396	0	149	4
1975	6646	0	6646	122	1796	22	1753	29	46	31	671	0	1987	0	184	4
1980	8857	0	8857	173	2261	30	2248	40	73	14	835	0	2950	0	228	4
1985	11502	0	11502	150	2709	36	2910	61	119	0	928	0	4254	0	331	4
2000	23876	0	23876	245	4771	55	5487	180	227	0	1613	0	10697	0	597	4
GAS																
1970	21255	846	22101	171	8884	281	1804	7	0	968	4015	595	4722	0	582	72
1975	22654	2279	24933	204	10215	347	2024	7	0	817	4245	599	5509	0	889	76
1980	23389	3550	26943	235	11389	414	1931	9	0	1231	4505	589	5716	0	869	55
1985	25276	3635	28882	278	13643	431	1742	13	0	1160	4856	568	5321	0	798	47
2000	28587	3932	32351	334	17425	450	1461	26	0	586	6472	475	4274	0	809	30

part of the forecast period, as an input to synthetic gas. The summary table brings out the fact that coal is entirely used as an intermediate fuel and that its use is projected to increase at an increasing rate.

Petroleum products are used heavily in all sectors. Here, growth rates are projected to differ sharply, for the demand for petroleum in final use increases only slowly, while intermediate demand continues to increase rapidly. Use of petroleum products in all intermediate sectors continues to increase at similar rates, with the average increase being 3.0 percent in the 1975 to 1985 period, then falling to 2.7 percent from 1985 to 2000. Final demand, particularly for personal consumption use, grows much more slowly, rising to a peak in the mid-1980s and then remaining steady over the rest of the forecast period.

The slow growth of final demand is due to various factors—higher petroleum prices and the conservation ethic's stimulating economy in use of petroleum, the incorporation of energy-saving devices such as more building insulation, heat pumps, and more efficient heating units into the capital stock of buildings, the replacing of large cars by more energy efficient vehicles, both smaller cars and more public transport, the substitution of electronic communication for some purposes presently achieved by moving people, and the increasing use of electricity in the home's partially replacing petroleum. The time involved in these changes is rather long, for

TABLE 12

COMPOSITION OF U.S. ENERGY USE, 1970-2000, SUMMARY

		USE (QUADRILLION BTU)			GROWTH RATES (AVERAGE PERCENT PER ANNUM)		
		FINAL	INTERMEDIATE	U S TOTAL	FINAL	INTERMEDIATE	U.S TOTAL
COAL	1970	0.368	12.554	12 922			
	1975	0 139	13 123	13 260	-17 69	0 89	0.52
	1980	0 111	14 193	14.304	- 4 40	1 58	1.52
	1985	0.113	16 424	16 538	0 36	2 96	2.94
	2000	0.092	26.799	26 891	- 1 36	3.32	3 29
PETROLEUM	1970	12.019	17 595	29 614			
	1975	13.694	20 002	33 696	2 64	2 60	2 62
	1980	16 045	23 131	39 128	3 22	2 95	3 06
	1985	17 545	27.002	44 547	1 80	3 14	2 60
	2000	17,864	40 284	58 148	0.12	2 70	1.79
ELECTRICITY	1970	1 545	3 673	5 218			
	1975	2 171	4.471	6 642	7 04	4 01	4 96
	1980	3 178	5 676	8.854	7 92	4 89	5.90
	1985	4 585	6 914	11 499	7 61	4 02	5 37
	2000	11 294	12.579	23 873	6 19	4 07	4 99
GAS	1970	5 304	16 725	22 029			
	1975	6 398	18 459	24 857	3.82	1 99	2.34
	1980	6 585	20 303	26.888	0 58	1 92	1 69
	1985	6 119	22.716	28 835	— 1 46	2 27	1 41
	2000	5 083	27.238	32 321	— 1 23	1 22	0 76

first attitudes must be changed, by economic, regulatory, and other pressures, and then these attitudes must be translated into new capital stock which, typically, must wait upon the depreciation and replacement of the existing stock, or in some cases, such as building insulation, can be incorporated into existing, highly durable stock. In any event, a long time lag is predicted before the full effects of this conservation become felt in terms of reduced energy input but ultimately, in the 1980s and 1990s, the scope for conservation is large enough to permit increasing levels of effective services to be sustained from a constant input of petroleum energy.

Electricity consumption increases at a rapid rate—5.3 percent a year between 1975 and 2000. The overall rate of increase steadily decreases, after an increase in the 1975 to 1980 period compared to the previous five years, due to the 1975 consumption figure's being depressed by the reaction to the current energy crisis. As in petroleum use, final and intermediate demand components grow at different rates but, with electricity, it is final demand that grows more rapidly. Intermediate use of electricity increases at an average of 4.2 percent a year with use in the services sector increasing the most rapidly, at 4.7 percent, and manufacturing use increasing at 4.0 percent. Final use of electricity increases at 6.8 percent a year with the most rapid increase taking place in personal consumption which increases at 6.9 percent a year over the 1975 to 2000 period. This use of electricity in personal consumption forms the most rapid increase of any use of any fuel and, in this, continues the same trend that has been observed in the past. The rate of increase is slower than that of the past due to the different price behavior, the new awareness in energy conservation, the introduction of more energy efficient capital such as heat pumps and thermal insulation, and the approach to saturation of such heavy electricity users as home air conditioning, dishwashers, laundries, ranges and so on.

Gas consumption is forecast to increase comparatively slowly and at a declining rate. Increasing use of gas is limited to two sectors—manufacturing and electricity generation—with use in other sectors, primarily services and personal consumption, declining. The net result is that final demand for gas declines, but intermediate demand continues to increase, although at a declining rate. The projected price increases for gas reduce its use in service and final use sectors, where its place is taken by petroleum and electricity, but gas remains important as an input into manufacturing and to electricity generation because of its nonprice advantages—it is clean burning, easily handled and it has many uses as a material, rather than a fuel, input.

The composition of inputs into the electricity generation sector is shown in Table 13. The dominant feature here is the projected increase in hydro and nuclear, virtually all the increase being in nuclear, generation. In 1970 hydro and nuclear provided 17.5 percent of total input but this proportion increases to 64 percent by 2000. Coal input, although growing rapidly in absolute terms, declines in relative importance as an input from supplying almost half of total inputs in 1970 to one-fifth in 2000. Similarly, although petroleum and gas inputs increase in absolute terms, this increase is relatively slow and they show a steady decline in proportion to total input.

TABLE 13

COMPOSITION OF INPUTS TO U.S ELECTRICITY GENERATION, 1970-2000 (PERCENT OF TOTAL INPUT)

	1970	1975	1980	1985	2000
COAL	45.4	41.5	37.3	33.2	21.6
PETROLEUM	12.7	11.5	9.4	7.9	5.3
GAS	24.4	20.6	16.4	13.6	8.8
HYDRO, NUCLEAR, OTHER	17.5	26.4	36.9	45.2	64.3

The changing composition of inputs into electricity generation reflects relative prices and is undertaken by the electricity generation sector in response to market forces. But, these forces also reflect, in part at least, a basic characteristic of U. S. fuel supply possibilities. Coal, uranium, and hydro power are not only relatively abundant but also have a low economic opportunity cost, for, apart from electricity absorbed in uranium enrichment, they have comparatively few alternative uses, whereas petroleum and gas fuels are not only in relatively short supply but also have many alternative uses that render their opportunity cost, for use in electricity generation, very high. In these respects, and omitting nuclear safety from consideration, the forecast outcome of market forces, as they affect inputs to electricity generation, appears to be entirely consistent with the objectives of currently proposed national energy policies.

☐ **Alternative projections.** This section presents summary comparisons between our energy projections and alternative projections. This comparison is conducted for two purposes: first, to obtain an independent check that our forecasts are of reasonable orders of magnitude and, second, to pinpoint the differences between economically based forecasts and technologically based forecasts. The alternative forecasts considered are those prepared by the Committee on the U. S. Energy Outlook of the National Petroleum Council,[26] referred to as NPC, the U. S. Department of the Interior forecasts prepared by Dupree and West, denoted DW, and, for electricity, various forecasts reported by Chapman, Mount, and Tyrrell,[27] denoted CMT.

The fossil-fuel projections of our interindustry model, denoted HJ, of the NPC, and of Dupree and West are shown in Table 14.

TABLE 14

ALTERNATIVE FORECASTS OF U.S. ENERGY UTILIZATION*

	DW		HJ		NPC	
	Q BTU	GROWTH RATE	Q BTU	GROWTH RATE	Q BTU	GROWTH RATE
COAL						
1970	12.922		12.922		12.922	
1975	13.825	1.4	13.260	0.5	16.1	4 5
1985	21.470	4.5	16.538	2.2	22.7	3.5
2000	31.360	2.6	26.891	3 3		
PETROLEUM						
1970	29.614		29.614		29.614	
1975	35.090	3 5	33.696	2.6	39.8	6.1
1985	50.700	3 7	44.547	2.8	56.2	3.5
2000	71.380	2.3	58.148	1.8		
GAS						
1970	22.029		22.029		22.029	
1975	25.220	2.7	24.727	2.3	21 3	-0.7
1985	28.390	1.2	28.145	1.3	21.1	-0.1
2000	33.980	1.2	28.681	0.1		
TOTAL INPUT						
1970	67.444		67 444			
1975	80.265	3.5	77 099	2.7		
1985	116.630	3.8	105 326	3.2		
2000	191.900	3.4	161.241	2.9		

* DW IS DUPREE AND WEST, HJ IS OUR PROJECTION, NPC IS THE NATIONAL PETROLEUM COUNCIL GROWTH RATES ARE AVERAGE ANNUAL PERCENTAGE RATES OF GROWTH. Q BTU IS QUADRILLION BTU.

For 1975 our forecasts and those of Dupree and West are similar, although for each fuel and for total input, ours are lower. The higher fuel prices incorporated into our projections reduce fuel use below the extrapolations of past trends used by Dupree and West. This same relationship continues for 1985 and 2000; i.e., our forecasts and those of Dupree and West are reasonably close, but the divergence increases over time, with our energy use always being below that in Dupree and West. This indicates first, that the orders of magnitude of our forecast are consistent with past energy trends used by Dupree and West, and second, that simple extrapolations of these trends are insufficient. Expected fuel price increases and changes in the nonenergy sections of the economy do exert a significant influence working to depress energy use below historical trends. The NPC forecasts for 1975 do not, at the present time, appear very realistic and the 1985 NPC forecasts are also open to question. The fact that our projections are well below those of NPC gives further weight to the need to take account of price and other factors that affect demand.

---

[26] See National Petroleum Council [36].

[27] See Chapman, Mount, and Tyrrell [1970]; a detailed report on the econometric model underlying these projections is given by Mount, Chapman, and Tyrrell [34].

Consider now the various projections of electricity consumption, given in Table 15. The striking impression gained from these forecasts is one of great variation—the projected 2000 consumption ranges from 2.0 to 10.3 trillion Kwh. Our forecasts are never the extreme entry but are always towards the lower end of the range of forecasts. To the extent that such a diversity of forecasts permits any conclusion, we can say that the above conclusions are reinforced. Our projections are in line with other forecasts but, by taking price and other economic factors into account, our projections are comparatively low.

TABLE 15

ALTERNATIVE FORECASTS OF U.S. ELECTRICITY CONSUMPTION (TRILLION KILOWATT HOURS)

PROJECTION*	1970	1975	1980	1985	1990	2000
CORNELL-NSF	1.57	2.15	2.92	3.96	5.38	10.25
CMT-HIGH	1.53	2.14	3.05	—	5.66	9.89
-MEDIUM	1.53	1.98	2.38	—	3.01	3.45
-LOW	1.53	1.88	2.07	—	2.11	2.01
DW	1.53	2.13	3.00	4.14	—	9.01
FPC	1.53	—	3.07	—	5.83	—
HJ	1.53	1.94	2.59	3.36	—	6.98
NPC	1.59	2.29	3.29	4.54	—	—

*CORNELL-NSF IS CORNELL-NATIONAL SCIENCE FOUNDATION WORKSHOP, CMT IS CHAPMAN, MOUNT, AND TYRRELL, DW IS DUPREE AND WEST, FPC IS FEDERAL POWER COMMISSION, HJ IS OUR PROJECTION, NPC IS NATIONAL PETROLEUM COUNCIL

SOURCES CORNELL-NSF, FPC, NPC FORECASTS ARE GIVEN IN CHAPMAN, MOUNT, AND TYRRELL (1972), TABLE 1, PAGE 3 THE CMT "LOW" FORECAST ASSUMES THAT THE REAL ELECTRICITY PRICE DOUBLES BY 2000, THE "MEDIUM" FORECAST CORRESPONDS TO THE FPC ESTIMATES OF A 19-PERCENT REAL PRICE INCREASE OVER 1970-1990, "HIGH" CORRESPONDS TO A 24-PERCENT REAL PRICE DECLINE OVER 1970-1980 AND A 12-PERCENT DECLINE OVER EACH OF THE 1980-1990 AND 1990-2000 INTERVALS

All the alternative energy forecasts considered here, apart from Chapman, Mount, and Tyrrell's electricity forecasts, have been prepared without explicit assumptions about the growth of real or relative fuel prices, without explicit consideration of interfuel substitutions and without explicit account of the interactions between energy and the rest of the economy or of future developments in the rest of the economy. This lack of analysis of the underlying forces at work in determining energy demand is a serious conceptual drawback. The omission is particularly significant in view of the present need to use energy models to aid the formation of future U. S. energy policy. Our interindustry model, however, avoids these problems and can provide a more realistic basis for policy analysis.

6. Energy policy: a Btu tax

Introduction. This section describes the application of the interindustry energy model to the analysis of one specific energy policy—a Btu tax designed to secure energy independence. This application serves to illustrate the methodology of the model and provides an evaluation of a specific tax proposal currently under consideration. The actual tax considered is a uniform rate of tax levied on the energy content of all fuels used outside the energy generation sector; such a tax is proposed in the Energy Revenue and Development Act of 1973, at present under consideration in the Senate Finance Committee.[28]

The starting point for the analysis of this tax is the base case set of energy projections which have been presented above. These projections are based on the assumption of no major new developments in energy policy and so can be used as a reference point against which the changes induced by policy changes are measured. In particular, the fuel imports in these forecasts define the objective of the policy, which is to reduce these imports to nominal levels by a target date, which we take to be 1985.

The tax has the effect of creating a wedge between the price paid by the energy consumer and that received by the producer. Since the forecast supply prices already equal average costs, the tax is translated into an increase in the fuel selling price, leaving the supply price unchanged, apart from indirect impacts on the supply price due to any production cost increases that are caused by the tax. Therefore, the tax leads to the reduction of fuel imports solely by acting on the demand side of the energy equation. This approach is the opposite of much of the present debate about energy policy, which is supply oriented. It emerges that demand based policies can be extremely effective and have the power to produce energy independence even without accompanying supply expansion policies.

The Btu tax is inserted into the model by means of price markups that increase the sales prices of fuels above the output price received by the seller. These markups vary for the five fuels since the Btu content of each dollar of fuel output is different for each fuel (the costs of Btu's obtained from the various fuels are later given in Table 18). The tax is assumed to be levied only on energy as it emerges from the fuel sectors so that energy inputs into fuel production, including the generation of electricity, are not subject to the tax. Also, exports of fuels, mainly coal, are considered exempt from the tax.

After price markups have been inserted, the model is solved to obtain a new set of economic and energy projections. This involves solving for new output prices and input-output coefficients, new final demand components, and the associated industry output levels and interindustry transactions. This information is then used to determine the energy deficits—the excess of U. S. demand for each fuel over the domestic output, both being calculated in terms of the new set of prices—which must be made up by fuel imports. This procedure captures both the direct and indirect effects of energy conservation—the direct saving of energy as an input by the substitution of other inputs for energy as well as the indirect saving of energy produced by substituting, in production and consumption, nonenergy intensive for the energy intensive goods and services.

---

[28] This section is based on Hudson and Jorgenson [24], presented as testimony at hearings by the Senate Finance Committee, January 16, 1974.

☐ **Tax policy and conservation.** We proceed in two steps, first examining the effects of the Btu tax by assessing the impact of various rates of tax on the energy and nonenergy sectors of the economy in 1980 and second, considering the specific tax structure that would be required to achieve the objective of independence from energy imports by 1985. The definition of energy independence used is for zero imports of gas and for only nominal imports of petroleum, such as would be required for fueling of U. S. aircraft and ships abroad. An arbitrary limit of one quadrillion Btu of energy imports was adopted for 1985; this limit is less than 10 percent of actual 1973 fuel imports.

The impact on energy use of various rates of Btu tax is summarized in Table 16. These figures show that the Btu tax can induce significant reductions in energy use and that it has the potential for reducing demand sufficiently to secure energy independence. The highest rate shown, $0.5 per million Btu, leads to a decline of 7 quadrillion Btu, or 7.8 percent in energy use relative to the no tax projection. This reduction is made up of substantial cuts in both final and intermediate uses of energy. The greater part of the reduction comes from the decline in energy input to intermediate production, but the relative cutback in use is greater in final demand where energy input is reduced by 11.8 percent, compared to the 6.6 percent fall in intermediate use.

TABLE 16

IMPACT OF BTU TAXES ON TOTAL ENERGY INPUT, 1980

TAX RATE ($/MILLION BTU)	0	0.1	0 3	0 5
TOTAL ENERGY INPUT (QUADRILLION BTU)	90 484	88.890	86.004	83.440
CHANGE FROM BASE		-1.593	-4.479	-7 043
CHANGE/TAX RATE		-15.930	-14.930	-14.070
PERCENT CHANGE FROM BASE		-1.760	-4.950	-7 780
INTERMEDIATE USE (QUADRILLION BTU)	70 356	69.321	67.415	65 692
CHANGE FROM BASE		-1.035	-2.941	-4 664
CHANGE/TAX RATE		-10 350	-9.800	-9.330
PERCENT CHANGE FROM BASE		-1.470	-4 180	-6 630
FINAL USE (QUADRILLION BTU)	20 127	19.569	18 589	17 748
CHANGE FROM BASE		-0 558	-1 538	-2 379
CHANGE/TAX RATE		-5.580	-5 130	-4 760
PERCENT CHANGE FROM BASE		-2.770	-7 640	-11.820
ENERGY IMPORTS (QUADRILLION BTU)	15.886	14.311	11.435	8.871
IMPORTS IN TOTAL INPUT (%)	17.600	16.100	13.300	10.600
TAX REVENUE ($ BILLION)	0	6.026	17 400	28.003

Our results indicate that although final users of energy may be more responsive to price increases than business users, the sheer volume of energy absorbed in production requires that, for maximum effect, energy conservation policies give at least as much weight to reducing intermediate as to reducing final demand. The response of both intermediate and final users to the tax varies with the tax level with, in both cases, the tax having a diminishing marginal impact on energy use. The decline in effectiveness is, however, gradual so that a reasonable first approximation is that each dollar of tax per million Btu reduces total energy input by about 15 quadrillion Btu. This corresponds to an elasticity of Btu use to Btu price of approximately —0.24.

The detailed adjustments by energy users to the imposition of the tax are shown in Table 17. These figures show, first, the different degrees of energy conservation in the four major energy consuming sectors—manufacturing, services, electricity generation, and personal consumption. Substantial economies are made in the energy input to each sector, but there is a wide difference in the proportionate response: energy used in personal consumption is reduced by 9.5 percent in response to the $0.5 tax rate, with services use being cut by 7.1 percent, manufacturing use by 5.8 percent, and input into electricity generation by 2.1 percent.

TABLE 17

IMPACT OF BTU TAXES ON INPUT PATTERNS, 1980

TAX RATE ($/MILLION BTU)	0	0.1	0.3	0.5
ENERGY INPUT, INCLUDING USE OF ELECTRICITY (QUADRILLION BTU)				
MANUFACTURING	18.908	18.669	18.222	17.909
SERVICES	10.496	10.335	10.032	9.752
ELECTRICITY GENERATION	27.403	27.289	27.063	26.840
PERSONAL CONSUMPTION	20.474	20.025	19.224	18.527
PERCENTAGE CHANGE IN TOTAL INPUTS				
ENERGY INPUT		-1.76	-4.95	-7.78
CAPITAL INPUT		0.11	0.29	0.49
LABOR INPUT		0.14	0.39	0.61
INPUT-OUTPUT COEFFICIENTS FOR TOTAL ENERGY INPUTS				
AGRICULTURE	0.0245	0.0239	0.0226	0.0216
MANUFACTURING	0.0235	0.0233	0.0232	0.0231
TRANSPORT	0.0447	0.0440	0.0426	0.0413
SERVICES	0.0190	0.0189	0.0186	0.0183
ELECTRICITY GENERATION	0.1864	0.1859	0.1850	0.1844
INPUT-OUTPUT COEFFICIENTS FOR MANUFACTURING				
COAL	0.0026	0.0026	0.0025	0.0025
PETROLEUM	0.0069	0.0068	0.0068	0.0067
ELECTRICITY	0.0085	0.0086	0.0087	0.0088
GAS	0.0055	0.0054	0.0052	0.0051
TOTAL ENERGY	0.0235	0.0233	0.0232	0.0231
CAPITAL SERVICES	0.1201	0.1200	0.1199	0.1198
LABOR SERVICES	0.2887	0.2889	0.2893	0.2897
MATERIALS INPUT	0.5164	0.5163	0.5162	0.5160

Energy use can be split into two broad categories—discretionary use and process use. Discretionary use includes inputs for comfort functions such as heating and cooling as well as personal services such as automobile travel. Process use covers fuel inputs for driving machinery, heating materials, turning generators, and so on. Discretionary uses are, typically, characterized by greater flexibility than process uses, firstly because the associated capital stock is generally less durable and easier to replace with more energy efficient capital, and secondly because discretionary uses are input to the generation of psychologically, rather than technically, specified performance, e.g., desired miles driven rather than energy inputs required per ton of alumina to be smelted, so the level of use within the existing capital stock can be more readily varied. Given this categorization, it is apparent that personal consumption includes a high ratio of discretionary to process energy use while this ratio for service activity is a little lower, for manufacturing lower still, and for electricity generation lowest of all. This ordering is precisely that observed in the sectoral response to the increase in energy prices.

More specifically, the input-output coefficients in Table 17 show the processes of adjustment to the tax induced increases in energy prices. Two types of adjustment can be seen. First, there is a substitution of the now relatively less expensive nonfuel inputs for the relatively more expensive fuel inputs. This is reflected in the total energy input coefficients which, for each sector, decline as the energy tax increases. The manufacturing input-output coefficients show the outcome of this substitution process: energy inputs decline, as do inputs of materials and inputs of capital, with inputs of labor services increasing to replace the three reduced inputs. Thus, in manufacturing, energy-capital-materials complementarity means that adjustments to economize on the expensive energy input lead to the reduction in the use of energy, capital, and materials and to the adoption of more labor-intensive production techniques.

The shift away from energy in manufacturing is relatively small in terms of changes in input-output coefficients, but it represents a substantial amount of energy. A similar process of substitution away from energy takes place in all sectors. The overall result, however, is a little different from that in manufacturing. The total reduction in energy use is made possible by an increase in inputs of both capital and labor. That is, on an economy-wide basis, energy-capital substitutability, such as the use of insulation or smaller cars to save energy, dominates the complementarity relation that characterizes manufacturing. The net effect is that the 7.8 percent fall in energy input is accommodated by a 0.6 percent increase in the demand for labor and a 0.5 percent increase in capital use and does not result in a comparable reduction in potential output.

The second set of adjustments occurs within energy input as interfuel substitution takes place. The tax has the effect of increasing different fuel prices by different extents, depending on the energy content of each fuel. Coal prices are increased to the greatest extent, with gas prices next, then petroleum, with electricity prices being increased the least. The input-output coefficients for fuel use in manufacturing, shown in Table 17, illustrate the resulting substitutions. Coal use declines only slightly for, despite the sharp price increase, coal remains the cheapest source of energy as well as being, for technical reasons, used in some production, such as steel, regardless of the price changes. The petroleum and gas coefficients both decline more noticeably. Electricity use, however, increases since electricity has become a relatively less expensive fuel and since its flexibility in use permits it to substitute for petroleum and gas.

The impact of the Btu tax on the price and consumption of each fuel is shown in Table 18. Coal prices increase by the largest proportion, with the average increase of 28 percent, for the $0.5 tax rate, comprising a 70-percent increase in price to nonfuel purchasers and virtually no increase in price in fuel production or export uses. The gas price increases by 24 percent for the highest tax rate, with the wholesale price of refined petroleum products increasing by 23 percent and electricity rising by only 7 percent. The small price increase for electricity leads to a correspondingly small demand decline, only 2 percent. This small decline also implies that coal input into electricity, the main use of coal, declines by only a small proportion. The only other sizeable domestic use of coal is in the manufacturing sector, but this use is relatively insensitive to price as discussed above. The price of gas increases substantially and this, with the opportunities to economize in its use and to substitute other fuels for it, leads to a decline of 12 percent in its use. Similarly, petroleum prices rise substantially, causing a decline of 8 percent in the consumption of petroleum. The average price elasticities implied by these consumption responses are: —0.08 for coal, —0.36 for petroleum, —0.31 for electricity, and —0.49 for gas.

TABLE 18

IMPACT OF BTU TAXES ON ENERGY CONSUMPTION AND PRICES, 1980

TAX RATE ($/MILLION BTU)	0	0.1	0.3	0 5
FUEL CONSUMPTION (QUADRILLION BTU)				
COAL	14.304	14 237	14.104	13 971
PETROLEUM	39 176	38.459	37.138	35 940
ELECTRICITY	8 854	8 817	8 744	8 672
GAS	26 888	26.126	24 791	23 650
CHANGE IN CONSUMPTION FROM BASE				
COAL		-0 47	-1 40	-2 33
PETROLEUM		-1.83	-5 20	-8 26
ELECTRICITY		-0 42	-1 24	-2 06
GAS		-2 83	-7 80	-12 04
AVERAGE FUEL PRICES				
COAL ($/TON)	16 3800	17.3100	19.1600	21 0200
REFINED PETROLEUM ($/BARREL, WHOLESALE)	11 8400	12.3800	13.4600	14 5500
ELECTRICITY ($/KWHR)	0 0256	0 0259	0.0266	0 0273
GAS ($/TH CU.FT)	1 4800	1 5500	1.7000	1 8400
CHANGE IN PRICES FROM BASE (PERCENT)				
COAL		5 68	16 94	28 33
REFINED PETROLEUM		4 57	13 71	22 85
ELECTRICITY		1 32	3 96	6 60
GAS		4.87	14 62	24 36
PRICE OF ENERGY TO TAXED USERS ($/MILLION BTU)				
COAL	0 71	0.81	1.01	1 21
PETROLEUM	2 21	2 31	2 51	2 71
ELECTRICITY	7 46	7 56	7 76	7 96
GAS	1 42	1.52	1 72	1 92

The effects of the Btu tax are not restricted to the energy sectors. There are also effects on the input structure of other production, as have been outlined above, on the costs and prices of these other goods, on the demand for these goods in intermediate uses, and on the level and composition of real final demand. These effects are summarized in Table 19, which gives the tax induced changes in nonenergy prices and quantities demanded. All prices are increased as the effects of the higher fuel prices work through the production cost structure. The process of substitution towards the relatively less expensive inputs lessens the impact of the fuel price increases on average costs, but the net effect is still an upward movement in costs, and in prices in general.

TABLE 19

IMPACT OF BTU TAXES ON NONENERGY PRICES AND QUANTITIES, 1980 PERCENTAGE CHANGE FROM BASE CASE

	PRICE			QUANTITY		
TAX RATE ($/MILLION BTU)	0.1	0.3	0.5	0.1	0.3	0.5
GROSS OUTPUT						
AGRICULTURE	0.17	0.51	0.82	-0.18	-0.53	-0.85
MANUFACTURING	0.16	0.50	0.82	-0.10	-0.30	-0.49
TRANSPORT	0.21	0.64	1.05	-0.17	-0.52	-0.84
SERVICES	0.06	0.22	0.34	-0.08	-0.22	-0.36
FINAL DEMAND						
CONSUMPTION	0.27	0.86	1.34	-0.12	-0.33	-0.53
INVESTMENT	0.16	0.48	0.74	-0.11	-0.31	-0.51
GOVERNMENT	0.12	0.38	0.56	-0.08	-0.22	-0.36
GNP	0.23	0.69	1.07	-0.09	-0.26	-0.42

The price increases by sector are in line with the importance of energy in the sector's inputs: transport prices increase the most, followed by agriculture and manufacturing, with service prices rising the least; this ranking is the same as the ranking of these sectors in terms of energy input-output coefficients. The price increases are not, however, very sizeable. Even for the $0.5 tax rate, the price increases range from 0.34 percent for services to 1.05 percent for transport. Similarly, the aggregate consumption price index and the GNP price deflator increase by only about 1 percent since the small nonenergy price rises dominate the larger fuel price increases in these price indices. The quantity changes induced by the energy tax are correspondingly small. The new prices, and the decline in real incomes, lead to a reduction and redirection of real final demand, but the resulting change in total real consumption and real GNP are very small, only of the order of 0.5 percent.

Two alternative bases for the Btu tax were examined. These were the cases in which the tax was levied only on final consumption of energy and in which the tax was levied only on intermediate inputs to the nonenergy sectors. Both taxes do reduce total energy input, the final use tax by 2.8 percent and the intermediate tax by 5.3 percent for a tax rate of $0.5 per million Btu. This compares with the 7.8 percent reduction for the general tax. Thus, although the energy conservation impact of a Btu tax diminishes as the base is restricted, even a tax on one of the narrow bases does lead to nonnegligible cutbacks in energy consumption. So, if political considerations were to preclude a final use tax, the intermediate Btu tax could still play an effective role in an energy policy package.

☐ **Import independence.** Our central conclusion is that a Btu tax can induce significant reductions in energy consumption, corresponding to even larger relative reductions in fuel imports, at the cost of comparatively minor changes in production patterns, prices, or demand. Our next step is to apply these results to formulate a specific Btu tax program that would result in import independence by 1985. The tax rate required to reduce energy consumption in 1985 to a level that requires no fuel imports is $1.35 per million Btu. The effects of this tax are summarized in Table 20. The revenue yield of this tax is $76.9 billion. This is a substantial withdrawal from the private spending stream and, unless returned through other policy measures, would have a significant deflationary impact. This revenue would be more than ample to sustain a major research and development effort on energy supply as well as to cover other government spending or to permit reductions in other taxes.

TABLE 20

IMPACT OF AN ENERGY INDEPENDENCE BTU TAX, 1985

TAX RATE ($/MILLION BTU)	0	1.35
TAX REVENUE ($ BILLION)	0	76.942
TOTAL ENERGY INPUT (QUADRILLION BTU)	105.326	88.244
PERCENTAGE CHANGE FROM BASE CASE		-16.2
IMPORTS OF PETROLEUM (QUADRILLION BTU)	9.628	0.975
IMPORTS OF GAS (QUADRILLION BTU)	3.635	0
IMPORTS IN TOTAL INPUT (PERCENT)	12.6	1 1
FUEL CONSUMPTION (QUADRILLION BTU)		
COAL	16.538	15.730
PETROLEUM	44.547	36.255
ELECTRICITY	11.499	10 780
GAS	28.835	21 926
PERCENTAGE CHANGE IN FUEL CONSUMPTION FROM BASE CASE		
COAL		-4 89
PETROLEUM		-18.61
ELECTRICITY		-6.25
GAS		-23 96
PERCENTAGE CHANGE IN PRICES FROM BASE CASE		
COAL		49.15
PETROLEUM		42.20
ELECTRICITY		15.57
GAS		47 16
AGRICULTURE		1.78
MANUFACTURING		1.70
TRANSPORT		1.98
SERVICES		0 71
CONSUMPTION		2.41
GNP		2.04

The reduction in energy use is secured through the effect of the tax in increasing fuel prices. In fact, a rise of around 40 percent in fuel prices, compared to the base case, is required to obtain the necessary cutback in energy use. Average coal prices are raised by 49 percent by the tax, gas prices are increased by 47 percent, petroleum prices by 42 percent, and electricity prices by 15 percent. The induced reductions in fuel consumption range from 5 percent for coal and 6 percent for electricity, to 19 percent for petroleum, and 24 percent for gas. The total reduction in energy input is 17.1 quadrillion Btu, which is 16.2 percent of the base case level. This reduction corresponds to a greater reduction in energy input per dollar of Btu tax rate than emerged from the 1980 simulations. The reason for this is the lags involved in responses to higher energy prices, particularly the lags involved in the replacement of capital stock with more energy efficient capital.

These reductions in fuel consumption are greater than are strictly necessary for energy independence. Coal and electricity suffer no supply deficit and reduction in the use of these fuels does not directly contribute to the independence objective. Gas use is cut back substantially by the tax, by much more, in fact, than the 10 percent reduction that would be required to eliminate gas imports. Thus, although the Btu tax does, fortuitously, induce conservation primarily in the two fuels whose consumption must be reduced if the independence objective is to be achieved, it is comparatively inefficient in achieving this objective, for it is so broad in its effects that the tax rate required to achieve petroleum independence produces excessively large reductions in the use of other fuels. The application of a general instrument in pursuit of such a specific objective results in unnecessary economic cost. Since excess petroleum demand is the binding constraint in the independence objective, a specific petroleum tax would be a more efficient instrument.

The effects of the energy tax on the nonfuel sectors are not large. Prices increase by an average of only 2 percent, for the large fuel price increases are dominated, in the consumption and GNP price indices, by the smaller increases in the prices of nonfuel goods. The input substitution processes, described above, permit production to accommodate the large reductions in energy input and the increases in fuel prices without corresponding decreases in output or increases in total costs.

The specific rate timetable of a Btu tax that would achieve energy independence by 1985 is a variable. The rate structure is subject to constraints, the rate must be zero in 1974 and $1.35 per million Btu in 1985, but the path between these end points is somewhat arbitrary. One possible rate timetable, with its associated macro-economic effects, is presented in Table 21. But, if the Btu tax program were to be implemented, its rate structure should satisfy certain additional conditions. First the tax must be regarded as a permanent measure, for, in the absence of other policies, continued energy independence requires that energy demand continually be depressed by taxes, and this will probably mean steadily increasing rates of tax. Second, the rate level should be increased fairly rapidly to the $1.35 level for, in view of the delays in incorporating energy conservation measures into production and final consumption, a high tax rate over much of the 1974 to 1985 period would not only induce substantial conservation but would induce it early enough

TABLE 21

A BTU TAX FOR ENERGY INDEPENDENCE IN 1985

	1975	1980	1985
BTU TAX RATE ($/MILLION BTU)	0	0.5	1.35
ENERGY INPUT (QUADRILLION BTU)	77.1	83.4	88.2
ENERGY IMPORTS (QUADRILLION BTU)	15.9	8.9	1.0
REVENUE FROM TAX ($ BILLION)	0	28.0	76.9
CONSUMPTION OF ENERGY (QUADRILLION BTU)			
COAL	13.3	14.0	15.7
PETROLEUM	33.7	35.9	36.3
ELECTRICITY	6.6	8.7	10.8
GAS	24.7	23.7	21.9
PERCENTAGE CHANGE FROM ZERO TAX CASE			
REAL CONSUMPTION	0	-0.53	-1.22
CONSUMPTION PRICE	0	1.34	2.41
REAL GNP	0	-0.42	-0.99
GNP PRICE DEFLATOR	0	1.07	2.04

for its full effects to be felt by 1985. Third, it is desirable that the rate structure be known for several years into the future so as to reduce any economic incentive for delaying energy conservation in the hope that energy prices would fall in the future.

Our results show that a Btu tax can induce sufficient fuel conservation to secure energy independence by the mid-1980s. Further, it suggests that the cost of achieving this is not catastrophic. But it does not follow that this is the best or only means for securing independence. Arguments could be made for a policy mix that stimulated supply expansion as well as demand reduction. Also, arguments could be made for a mix of policies on the demand side, such as a Btu tax, regulations requiring building insulation, taxing of particularly heavy energy uses, and so on. And it might be argued that a general policy, such as a Btu tax, is inefficient, for it reduces demand for all fuels when the critical problem is evidenced only in petroleum. More fundamentally, the desirability and definition of energy independence could be debated *ad infinitum.* Our results beg the question of underlying objectives, but they do indicate that the Btu tax does provide a means of securing energy independence.

7. Conclusion

■ We conclude with a brief overview of our methodology, comparing our framework for the analysis of energy policy with alternative approaches. Our model replicates each of the components of the overall transactions flow—purchase of primary inputs, sales of goods and services between sectors, formation of product prices, and purchase of output by final users. These aspects of economic activity are brought into consistency by means of simulated market processes. All decision units react to the same set of prices, and prices and quantities adjust so that all markets are cleared and each production sector covers its costs. The heart of the model lies in a series of submodels of production behavior, one for each of the nine domestic producing sectors. This set of production relationships provides the basic information used to determine relative output prices and the corresponding set of input patterns.

The sectoral production models are based upon price possibility frontiers. Within each sector, the price frontier expresses output price in terms of the prices of inputs and the production efficiency of that sector. Each frontier contains first- and second-order terms so that both the average requirements for each input and the interrelationships among inputs are captured. Thus, although the specification is in terms of prices, it captures the same information concerning input requirements, complementarities, and substitutabilities as the more traditional production function.

The primary input prices are generated within our macro-econometric growth model. Productivities are exogenous to the producer submodels so the nine sectoral output prices can be solved from the nine price frontiers. The simultaneous determination of prices permits the derivation of a set of prices for all produced output which not only takes account of production constraints and interrelationships and primary input prices, but also integrates all these price determinants into a consistent framework so that all sectors are simultaneously charging the minimum price that covers all their costs, including a return to capital. In short, given primary input prices and underlying production information, the model finds the equilibrium price system for the economy.

This approach to prices and production also permits the input pattern for each sector that is best suited to the prevailing set of prices and costs to be chosen from the infinity of patterns that are possible. Thus, producers can react to the prevailing structure of relative prices by adjusting their input patterns and substituting, within the limits set by the complementarity and substitutability information contained in the price frontiers, relatively less expensive for relatively more expensive inputs. In this way, a critical feature of the economy, the response of input patterns to prices, is systematically included in our model. This feature of the model is very important from a practical point of view in accommodating the detailed operation of the price system. Conceptually, it frees input-output analysis from the assumption of fixed coefficients, and extends input-output analysis by making the coefficients variable and by making them fully endogenous and linking them to price behavior.

After the analysis of production relationships has been completed, the next step in implementing our model is to examine the components of final demand. These components are produced by three final demand submodels. The initial inputs into these submodels are the prices charged for the output of each of the domestic producing sectors, the prices of primary inputs that enter directly into final use, and the total current dollar expenditures in three of the final demand categories—personal consumption, private investment, and government purchases. These expenditure totals are produced by the use of our macro-econometric model of U. S. economic growth. Our growth model provides levels of relative prices of consumption, investment, capital, and labor within the framework employed to generate final demand in real terms. The model represents a major extension of macro-econometric model building to incorporate elements of both aggregate demand and aggregate supply.

We begin with the set of prices that is consistent with equilibrium demand and supply and with the levels of prices of primary inputs. We find the pattern of input-output coefficients associated with this price regime. We determine the level and composition of final demand that are associated with these prices. The input-output coefficients have been determined already, so the balance equations of input-output analysis are sufficient to find, in constant dollar terms, the total sectoral outputs and the pattern of industry purchases and sales necessary to sustain the final demand. This step imposes a final set of market clearing relationships on our simulated economy with the condition of equality between the real demand and supply of every commodity's being added to the condition of equality between value of demand and supply and of equality between receipts and expenditures for every sector.

The output from the interindustry model comprises prices for each of the supplying sectors, the matrix of constant dollar transactions for the entire structure, and, by combining these two elements, the matrix of current dollar transactions. The transactions matrices cover the whole economic structure—energy as well as nonenergy sectors. The energy information contained in these transactions matrices is already extensive, covering both volumes and prices. It can readily be extended, however, to produce the data forms traditionally used in energy analysis—energy flows in British thermal units, energy flows in physical units, and energy prices in terms of physical units. This further information for Btu's can be generated by inserting known values of the Btu content per constant dollar of each fuel, the volume in physical units can be obtained by using known physical units per constant dollar ratios, and fuel prices can be generated by applying the base period prices to the price indices simulated in the model.

We conclude this overview of our framework for energy policy analysis by comparing the conceptual properties of our model with those of the traditional energy balance approach to energy analysis. The fundamental advantage of our approach is that, rather than viewing energy in isolation, we view it as one of the many interacting parts that make up the economic system. This perspective permits the systematic analysis of all the factors that influence energy on both demand and supply sides and, equally important, it permits the explicit linkage of energy developments to those variables such as employment, incomes, and consumption, that are the ultimate ends to which energy use is only a means.

More specifically, our model incorporates the influence of fuel prices on the level and composition of energy use and, further, it incorporates the effects of the level and pattern of nonenergy activity on energy use as well as the reverse linkage of energy prices and supplies to nonenergy price input, output, and consumption patterns. These interrelationships are critical and it is essential, for both forecasting and policy purposes, to recognize them. Some examples of these linkages are the severe impact of the recent oil shortages on the output of and the incomes generated by the automobile and tourist industries, the implications for energy use of the secular trend of demand towards service activities, and the economies in fuel use induced by the recent increases in fuel prices.

One aspect of these linkages in which our model represents a particularly important advance over energy balance procedures is the variation of input patterns in response to relative prices. This is of great importance in energy analysis in view of the widespread ability to substitute between fuels, for example between coal and oil in electricity generation, between oil and gas in industrial heating, between gas and electricity in home cooking, and so on. This substitution is not merely a matter of switching fuels, for the associated capital stock must also be changed to permit the use of a different fuel. Our production models, by treating all inputs simultaneously, take account of the possibilities for both interfuel substitutions and substitutions of fuel for nonfuel input, but do this in such a way that the constraints implied by complementarities between inputs are recognized, so that a consistent analysis of the entire input picture is obtained.

Our investigation of the effects of a Btu tax serves to demonstrate the usefulness of our framework for energy policy analysis. The basic properties of the model are illustrated by the result that, in the 1980 simulations for example, energy input can be reduced by 8 percent at the cost of only a 1-percent increase in average prices and a 0.4-percent decrease in real income. In other words, the flexibility of the economy in adapting to changing resource availabilities, and the power of the price system in securing this adaptation, mean that substantial reductions in energy use can be achieved without major economic cost. The analytical property of our model that incorporates this flexibility is the endogenous formation of prices and the endogenous determination of the response of production patterns and final demand to these prices. Also the integration of the various components of the model by means of interindustry analysis secures an overall consistency in the simulation of the market process. These features, permitting price formation and the reaction of producers and consumers to price changes, combined in a simultaneous model of the entire economy, represent a major advance over traditional interindustry and energy balance analysis.

References

1. Almon, C., Buckler, M R., Horwitz, L. M., and Reimbold, T. C. *1985: Interindustry Forecasts of the American Economy*. Lexington, Mass.: Heath, 1974.
2. Berndt, E. R., and Christensen, L. R. "The Specification of Technology in U. S. Manufacturing." Discussion Paper 73-17, University of British Columbia, November, 1973.
3. ———. "Testing for the Existence of a Consistent Aggregate Index of Labor Inputs." *The American Economic Review,* Vol. 64, No. 3 (June 1974), pp. 391–404.
4. ———. "The Translog Function and Substitution of Equipment, Structures, and Labor in U. S. Manufacturing, 1929–1968." *Journal of Econometrics,* Vol. 1, No. 1 (March 1973), pp. 81–114.
5. Berndt, E. R. and Jorgenson, D. W. "Production Structure," Chapter 3 in Jorgenson *et al.* [26].
6. Berndt, E R. and Wood, D. W. "Technology, Prices, and the Derived Demand for Energy." Discussion Paper 74-09, University of British Columbia, May 1974.

7. Burmeister, E. and Dobell, A. R. *Mathematical Theories of Economic Growth,* New York: Macmillan, 1970.
8. Burrows, J. C. and Domencich, T. *An Analysis of the United States Oil Import Quota.* Lexington, Mass.: Heath-Lexington, 1970.
9. Carter, A P. *Structural Change in the American Economy.* Cambridge: Harvard Univ. Press, 1970.
10. ——— and Brody, A. *Contributions to Input-Output Analysis.* Amsterdam: North-Holland, 1970.
11. Chapman, D., Mount, T. and Tyrrell, T. "Electricity Growth: Implications for Research and Development." Testimony before the Committee on Science for Astronautics, U. S. House of Representatives, June 16, 1972.
12. Christensen, L. R. and Jorgenson, D. W. "Intertemporal Optimization and the Explanation of Consumer Behavior." Paper presented at the Winter Meetings of the Econometric Society, December, 1968.
13. ———. "The Measurement of U. S. Real Capital Input, 1929–1967." *Review of Income and Wealth,* Series 15, No. 4 (December 1969), pp. 293–320.
14. ———. "Measuring Economic Performance in the Private Sector," in M. Moss, ed., *The Measurement of Social and Economic Performance,* Studies in Income and Wealth, No. 37, New York: Columbia Univ. Press, 1973, pp. 233–351.
15. ———. "U. S. Income, Saving, and Wealth, 1929–1969." *Review of Income and Wealth,* Series 19, No. 4 (December 1973), pp. 329–362.
16. ———. "U. S. Real Product and Real Factor Input, 1929–1967." *Review of Income and Wealth,* Series 16, No 1 (March 1970), pp. 19–50.
17. Christensen, L. R., Jorgenson, D. W., and Lau, L. J. "Conjugate Duality and the Transcendental Logarithmic Utility Function." *Econometrica,* Vol. 39, No. 4 (July 1971), pp. 255–256.
18. ———. "Transcendental Logarithmic Production Frontiers." *Review of Economics and Statistics,* Vol. 55, No. 1 (February 1973), pp. 28–45.
19. ———. "Transcendental Logarithmic Utility Functions." *The American Economic Review,* forthcoming.
20. Dupree, W. and West, R. *United States Energy through 2000.* U. S. Department of the Interior, Washington, D. C., December, 1972.
21. Faucett, J. and Associates. *Data Development for the I-O Energy Model,* Final Report to the Energy Policy Project, Washington, D. C., May, 1973.
22. Goldberger, A. S. *Impact Multipliers and Dynamic Properties of the Klein-Goldberger Model.* Amsterdam: North-Holland, 1959.
23. Hickman, B., ed. *Econometric Models of Cyclical Behavior.* Studies in Income and Wealth, No. 36. New York: Columbia Univ. Press, 1972.
24. Hudson, E. A. and Jorgenson, D. W. "Tax Policy and Energy Conservation." Testimony before the Committee on Finance, U. S. Senate, January 16, 1974.
25. ———. "U. S. Economic Growth, 1973–2000," in Data Resources, Inc., *Long Term Projections of the U. S. Economy,* Lexington, Mass., April, 1974, Chapter 17, pp. 161–236.
26. Jorgenson, D. W., Berndt, E. R., Christensen, L. R. and Hudson, E. A. *U. S. Energy Resources and Economic Growth,* Final Report to the Energy Policy Project, Washington, D. C., September, 1973.
27. Klein, L. R. and Goldberger, A. S. *An Econometric Model of the United States, 1929–1952.* Amsterdam: North-Holland, 1955.
28. Lau, L. J. "Econometrics of Monotonicity, Convexity, and Quasiconvexity." Technical Report No. 123, Institute for Mathematical Studies in the Social Sciences, Stanford University, March, 1974.
29. Leontief, W. W. "Introduction to a Theory of the Internal Structure of Functional Relationships." *Econometrica,* Vol. 157, No. 4 (October 1947), pp. 361–373.
30. ———. *The Structure of the American Economy 1919–1939.* New York: Oxford Univ. Press, 1951.
31. ———. *Studies in the Structure of the American Economy.* New York: Oxford Univ. Press, 1953.

32. MALINVAUD, E. *Statistical Methods of Econometrics.* Amsterdam: North-Holland, 1970.
33. MORISHIMA, M. "A Dynamic Leontief System with Neo-Classical Production Functions." *Equilibrium Stability and Growth.* Oxford: Oxford Univ. Press, 1964, pp. 54–92.
34. MOUNT, T. D., CHAPMAN, L. D. AND TYRRELL, T. J. "Electricity Demand in the United States." Oak Ridge National Laboratory, June, 1973.
35. NATIONAL PETROLEUM COUNCIL. *U. S. Energy Outlook: An Initial Appraisal 1971–1985.* Washington, D. C., July, 1971.
36. ———. *U. S. Energy Outlook.* Washington, D. C., December, 1972.
37. SAMUELSON, P. A. "Nonsubstitution Theorems," in J. Stiglitz, ed., *The Collected Scientific Papers of Paul A. Samuelson,* Cambridge: M.I.T. Press, Vol. 1, pp. 513–536.
38. SUITS, D. B. "Forecasting and Analysis with an Econometric Model." *The American Economic Review,* Vol. 52, No. 1 (March 1962), pp. 104–132.
39. U. S. FEDERAL POWER COMMISSION. *The 1970 National Power Survey.* Washington, D. C.: December, 1971.
40. VACCARA, B. N. "Changes over Time in Input-Output Coefficients for the United States," in Carter and Brody, eds., [10], pp. 238–260.
41. WALRAS, L. *Elements of Pure Economics,* translated by W. Jaffé. Homewood, Ill.: Irwin, 1954.

The effects of higher prices on electricity consumption

James M. Griffin

Associate Professor of Economics
University of Pennsylvania

This paper utilizes an econometric model of the electric utility industry to simulate the effects of higher fuel prices on electricity demand and on the mix of fuel inputs to electricity generation. The model treats as endogenous electricity demand, electricity prices, the efficiency of fuel conversion, and the choice among coal, natural gas, residual fuel oil, and nuclear fuel inputs. The results here suggest that given projected fuel input prices, the short- and long-run impacts on electricity demand are likely to be small. The effects on fuel mix appear quite substantial in the intermediate and long run, but in the period of one year or less, fuel substitution possibilities appear fairly limited.

1. Introduction

■ The current debacle called the "energy crisis" could be more aptly described as an "energy *policy* crisis" to characterize the gravity and complexity of choosing the proper set of policy instruments. Up until recently, research emphasis has centered on estimating the elasticity of supply for the primary energy sources, particularly crude oil and natural gas. While these questions are of critical importance to long-run policy, the recent shortages of petroleum products underscore the short-term significance of demand for secondary energy forms and potentials for interfuel substitution. For example, the viability of proposals to tax various energy forms such as gasoline depends largely on the magnitude of dynamic price elasticities of demand. Similarly, the success of proposals to alter air pollution controls to alleviate fuel shortages depends on the degree of short and intermediate term substitution possibilities among fuels in the production of secondary energy forms such as electricity.

The electric utility industry provides interesting examples of both phenomena. Consumers of electricity can presumably substitute other goods, which may include other energy or nonenergy goods, for electricity in response to changes in relative prices and technology. In contrast to other energy products such as motor

---

James M. Griffin received the A.B. degree in economics and mathematics from Southern Methodist University (1966) and the Ph.D. in economics from the University of Pennsylvania (1970). His research interests include energy modeling and policy issues in natural resource allocation.

The author wishes to acknowledge the numerous helpful suggestions by F. G. Adams and Wade P. Sewell. This study was partially financed by a faculty development research grant by the University of Houston and the Ford Foundation's Energy Policy Project.

gasoline, there also exists the possibility of fuel substitution in the production of electricity. For example, depending on the type of generation facilities, electricity can be generated by hydroelectric power, nuclear power, coal, natural gas, and residual fuel oil. Prior to the construction of a plant, interfuel substitution is clearly possible, but even within existing plant capacity there is a limited capability of substitution among the fossil fuels. In 1969, it was estimated that approximately 8.3 percent of steam electric generating stations could use all three fossil fuels.[1] In the interim, air pollution regulations have resulted in greater flexibility as many coal burning plants have been converted to use petroleum and yet have maintained the coal facilities intact.

The policy implications of rapidly escalating fuel prices both on the demand for electricity and on fuel input substitution possibilities are the primary focus of this paper. For pedagogical purposes, the problem can be dichotomized in the following manner. First, rising prices of fuel and capital inputs to electricity generation can be expected to raise generation costs and electricity prices. It is important to be able to quantify this impact on electricity prices and their effects in turn on electricity demand. Second, since all fuel input prices are increasing at varying rates, the changing relative prices among fuels can be expected to alter substantially the mix of fuels consumed. The substitution possibilities among fuel inputs to electricity generation over the remainder of the decade are studied.

Given the current shortages of petroleum products, it is essential that we gain some idea of the potential for alleviating the problem through price adjustments. In sum, the demand for a particular fuel in electricity generation may be reduced as a consequence of decreased final demand for electricity and/or a change in the mix of fuel inputs.

In order to investigate final demand effects, it is necessary to consider the determinants of the sustained rapid growth in electricity output, which has expanded at a 7.2 percent annual rate since World War II and now accounts for 24 percent of the U.S. total energy requirements. It is vital to distinguish the relative effects of income, price, and technological factors because they are likely to differ significantly from the past. Utilizing a cross-section, time-series analysis, Fisher and Kaysen found that demographic and technological factors were primary determinants of household electricity demand and that price had only a negligible effect on long-term demand.[2] In contrast, a study by Wilson based on cross-section data seriously questions the finding of Fisher and Kaysen, concluding that "price is the major determinant of intermarket variations in the average volume of residential electricity consumption."[3] This paper seeks to resolve the controversy by estimating distributed lags on time series data to distinguish between short- and long-run income and price elasticities.

Analysis of the determinants of fuel inputs for electricity generation poses a difficult problem in itself. As indicated in Table 1,

---

[1] See Duchesneau [5].

[2] See [7].

[3] See [18].

TABLE 1

PERCENTAGE OF ELECTRICITY GENERATED BY ENERGY SOURCE[a]

	HYDRO-ELECTRIC	COAL	NATURAL GAS	FUEL OIL	NUCLEAR
1950	29.2	46.5	14.0	10.2	–
1960	19.3	53.4	21.0	6.1	0.1
1970	16.2	46.3	24.1	11.8	1.7

[a]SOURCE EDISON ELECTRIC [6]

the shares of the various energy inputs have changed significantly over time. In part, these mix changes can be ascribed to technical and environmental factors. No doubt, the Clean Air Act of 1970 and related provisions have had a pervasive effect on the demand for coal and low-sulphur fuel oil. Evidence also suggests that relative prices affect fuel choice both *ex post* and *ex ante.* MacAvoy found the share of new nuclear capacity to be affected by relative prices of nuclear fuel to fossil fuel and the relative price of nuclear capital to fossil capital costs.[4] Taking capacity as given, Verleger has demonstrated the significance of relative fuel prices in affecting the utilization of generating capacity.[5] For policy purposes, it is important to be able to describe the dynamic nature of these price effects on substitution possibilities over time.

2. An econometric model of the electric utility industry

General description. In order to assess the short- and long-run impacts of higher fuel prices on fuel requirements, an econometric model of the electric utility industry was developed with emphasis on its use *as a simulation tool.* The model attempts to characterize the determinants and interrelations of electricity demands, fuel inputs, and electricity prices by using a simultaneous equations system. The model is best suited for simulation and forecasting exercises over the intermediate term (five years). Use of it for longer term simulations is feasible, provided one is aware of its limitations.

The model is an annual model estimated with time series observations over the post war period for the aggregate United States. It consists of twenty-five equations, of which twelve are stochastic. There are ten behavioral equations, four linking equations, and eleven identities. The structure of the model is illustrated in simplified form in Figure 1. Each rectangle represents an endogenous variable, and the variables denoted by an asterisk at the bottom of the figure are exogenous. The arrows denote the flow of causation which runs from electricity demand to fuel conversion efficiency, then to fossil-fuel inputs, and finally to electricity prices. The model is not recursive, however, as electricity prices have a feedback effect on electricity demand. The sixteen exogenous variables to the model consist primarily of the prices of factor inputs (such as the various fuel input prices and capital costs) and mea-

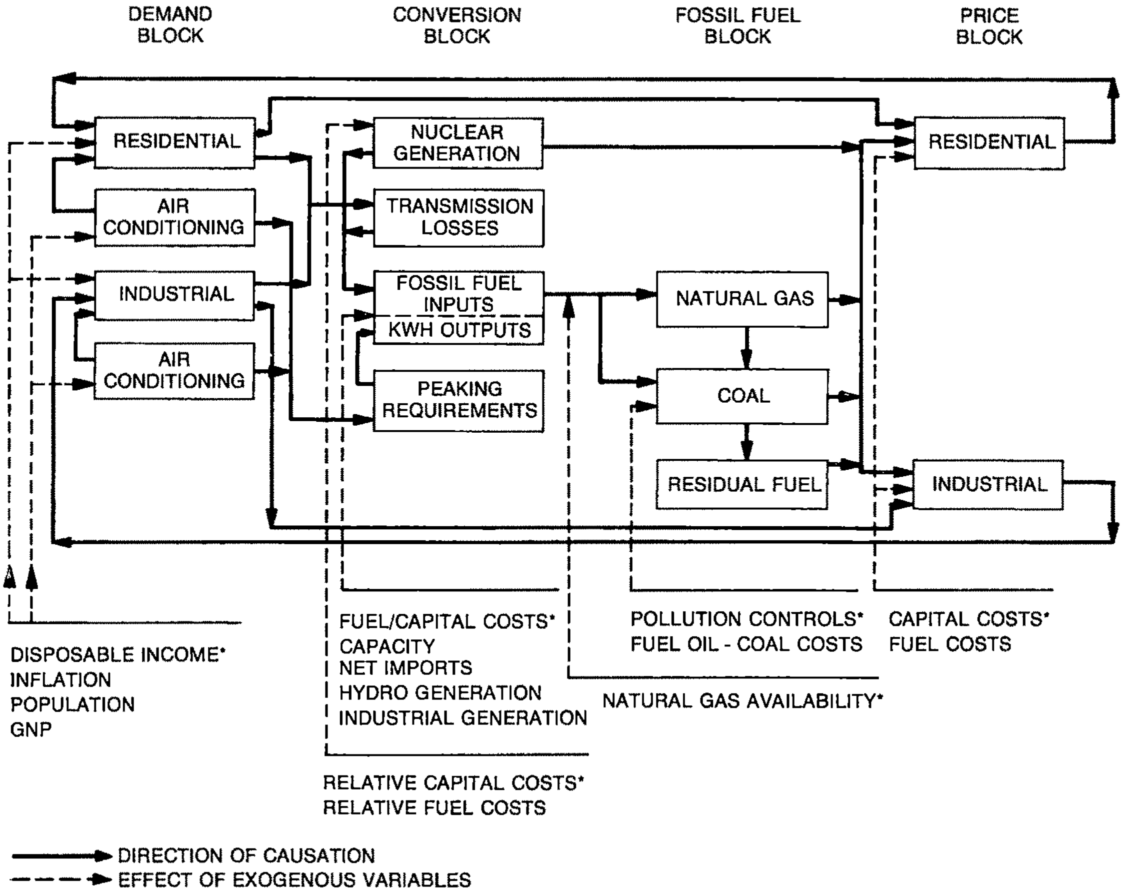
---

[4] See [9].

[5] See [16].

FIGURE 1

SIMPLIFIED STRUCTURE OF ELECTRICITY MODEL



sures of macroeconomic activity (such as real GNP, disposable income, and the GNP price deflator).

The development of microeconomic models for more than short-term forecasting applications presents at least two additional complications involving exogenous inputs and the estimation of distributed lags. One of the strengths of this study is that the electricity model is linked to the Wharton Long-Term and Industry Model,[6] which provides a macroeconomic setting within which to simulate the model. The use of the macroeconomic model as a source for inputs of real GNP, real disposable income, the GNP deflator, and the user cost of capital in utilities provides a fundamental compatibility among such variables. Moreover, effects of changes in interest rates and depreciation guidelines can be taken from the macro model and related to the industry through the user cost of capital variable.

In intermediate and long-term models, it is necessary to elicit the structure of the lagged relationships. The standard lagged dependent variable formulation imposes a common geometrically distributed lag on the independent variables, even though there may be little *a priori* basis for the imposition of such a lag structure. Also, ordinary least squares (OLS) estimates are generally inconsistent if

---

[6] See [12].

the disturbances are autocorrelated. For these reasons, this popular estimation approach in short-term models is rejected here in favor of the use of the Almon polynomial distributed lag procedure,[7] which places fewer constraints on the lag structure and reduces the problem of autocorrelation to an efficiency question. In addition, with a relatively small number of time series observations as in this study, the Almon procedure conserves degrees of freedom and reduces problems of multicollinearity.

The principal estimation approach is two stage least squares (2SLS) where the explanatory variables include a current period endogenous variable. It should be remarked that the model can almost be characterized as block recursive,[8] except for the minor effect of the current period's electricity price on demand. Dhrymes has shown that if the disturbances between equations are uncorrelated, ordinary least squares will yield consistent estimators.[9] In fact, earlier estimates using ordinary least squares provided quite similar estimates to those reported here.

A standard 5 percent *t*-test criterion was used for inclusion of variables which were first determined as relevant on theoretical grounds. In estimating the polynomial distributed lags, the lag structure which maximized $\bar{R}^2$, the coefficient of determination adjusted for degrees of freedom, was generally selected.

☐ **The model and its specification.** Table 2 provides a complete equation list showing the coefficient estimates, the *t*-statistics in parentheses, the adjusted coefficient of determination ($\bar{R}^2$), the equation's standard error (*S.E.*), the Durbin-Watson Statistic (*D.W.*), the estimation approach and the sample period. The order and degree of the polynomial lags are given in the footnotes. The variable names are identified both in the following discussion of individual equations and in Appendix 1.

The demand block:

The residential demand for electricity is in effect a derived demand depending on the stock of electricity-consuming equipment. In equation (1)*, residential demand per capita $\left(\frac{DR}{N}\right)$ is hypothesized to be a function of the per capita capital stock of the various electricity-consuming appliances $\frac{K_1}{N} \dots \frac{K_j}{N}$, and a random error term (e):

$$\frac{DR}{N} = f\left(\frac{K_1}{N}, \dots \frac{K_j}{N}\right) + e. \qquad (1)^*$$

In equation (2)*, the per capita stock of any given appliance

---

[7] See [1].

[8] For example, the equations can be ordered in the following manner:

$$\begin{aligned} y_1 &= \beta_1 X + e_1 \\ y_2 &= \beta_2 X + \alpha_1 y_1 + e_2 \\ &\cdots\cdots \\ y_n &= \beta_\lambda X + \gamma_1 y_1 + \gamma_2 y_2 \cdots \gamma_{n-1} y_{n-1} + e_n. \end{aligned}$$

[9] See Dhrymes [4], pp. 308–311.

TABLE 2

DESCRIPTION OF ELECTRICITY MODEL [a]

DEMAND BLOCK (1) $\frac{DR}{N} = \underset{(1\ 7)^{*}}{0\ 6486} + \underset{(6\ 6)}{0.00903}\,\frac{K1}{N} + \sum_{i=0}^{m}\beta_i\frac{YD^{b}}{N_{-i}} + \sum_{i=0}^{n}\Psi_i\frac{PR^{c}}{P_{-i}}$	$\beta_0 = 0\ 0287\ (0\ 9)^{*}$ $\beta_1 = 0\ 0686\ (2\ 3)$ $\beta_2 = 0\ 0937\ (2\ 7)$ $\beta_3 = 0\ 1043\ (2\ 8)$ $\beta_4 = 0\ 1002\ (2\ 7)$ $\beta_5 = 0\ 0864\ (2\ 7)$ $\beta_6 = 0\ 0480\ (2\ 7)$	$\Psi_0 = -0\ 0383\ (1\ 7)$ $\Psi_1 = -0\ 0612\ (2\ 0)$ $\Psi_2 = -0\ 0688\ (3\ 2)$ $\Psi_3 = -0\ 0609\ (11\ 2)$ $\Psi_4 = -0\ 0376\ (1\ 9)$
(2) $\frac{K1}{N} = \underset{(24\ 3)}{-159\ 7} + \sum_{i=0}^{m}\beta_i\frac{YD^{b}}{N_{-i}}$	$\beta_0 = 3\ 91\ (0\ 3)$ $\beta_1 = 20\ 35\ (11\ 3)$ $\beta_2 = 28\ 18\ (6\ 6)$ $\beta_3 = 27\ 34\ (4\ 2)$ $\beta_4 = 18\ 00\ (3\ 5)$	
(3) $\frac{DI}{X} = \underset{(7\ 3)}{1\ 6295} + \underset{(3\ 3)}{0\ 0164}\,\frac{KCA}{X} + \sum_{i=0}^{m}\beta_i\frac{PI^{b}}{P_{-i}}$	$\beta_0 = -0\ 0287\ (0\ 6)$ $\beta_1 = -0\ 0415\ (0.8)$ $\beta_2 = -0\ 0690\ (2\ 5)$ $\beta_3 = -0\ 0854\ (5\ 7)$	$\beta_4 = -0\ 0906\ (5\ 6)$ $\beta_5 = -0\ 0847\ (4\ 3)$ $\beta_6 = -0\ 0676\ (3\ 6)$ $\beta_7 = -0\ 0394\ (3\ 2)$
(4) $KCA = \underset{(25\ 0)}{-12472\ 5} + \sum_{i=0}^{m}\beta_i\ X^{d}_{-i}$	$\beta_0 = 4\ 07\ (39\ 1)$ $\beta_1 = 6\ 79\ (39\ 1)$ $\beta_2 = 8\ 14\ (39\ 1)$ $\beta_3 = 8\ 14\ (39\ 1)$ $\beta_4 = 6\ 78\ (39\ 1)$ $\beta_5 = 4\ 07\ (39\ 1)$	
CONVERSION EFFICIENCY BLOCK (5) $Z2 = \underset{(8\ 7)}{29\ 0} + \underset{(21\ 4)}{0\ 0743}\,(TG\text{-}Z3)$		
(6) $TG = DR + DI - Z1 + Z2$		
(7) $\log\frac{\Delta CN}{\Delta CN + \Delta C} = \underset{(9\ 6)}{-1\ 880} - \underset{(6\ 1)}{1\ 404}\log\frac{PNF}{PFF_{-7}} - \underset{(2\ 7)}{0\ 857}\log\frac{PNK}{PFK_{-7}}$		
(8) $Z5 = Z5_{-1} + \Delta(TG - Z3 - Z4)\ \frac{\Delta CN}{\Delta CN + \Delta C}$		
(9) $TGFF = TG - Z3 - Z4 - Z5$		
(10) $HR = \frac{TELBTU}{TGFF}\,10^{3}$		
(11) $HR - \frac{C_{-1}}{C}HR_{-1} = \underset{(11\ 3)}{6208\ 8}\,\frac{\Delta C}{C} + \underset{(5\ 9)}{3259}\,\frac{UCCP}{PFUEL}\,\frac{\Delta C}{C}$		
(12) $\frac{\Delta C}{C} = \frac{\Delta TGFF}{TGFF}$		
(13) $\left(\frac{Q_p}{Q_b} - 1.\right)(TG - Z3) = \underset{(0\ 0)}{0\ 2181} + \underset{(14\ 7)}{0\ 00503}\,(K1 + KCA)$		
(14) $UCCP = UC\left[1 + 6\left(\frac{Q_p}{Q_b} - 1\right)\right]$		

TABLE 2 CONTINUED

$\bar{R}^2$	S E	D W	ESTIMATION TECHNIQUE	SAMPLE PERIOD
0 996	0 0108	1 91	2SLS	1951-1971
0 984	3 43	0 48	OLS	1952-1971
0 990	0 0198	1 34	2SLS	1951-1971
0.988	463 6	0 44	OLS	1952-1971
0 958	5 93	0 66	OLS	1951-1971
0 726	0 559	—	OLS	21 OBS
0 896	57 7	1 40	2SLS	1952-1970
0 919	14 64	1 89	OLS	1952-1971

TABLE 2 CONTINUED

FOSSIL FUEL BLOCK		
(15) GELBTU = −705 7 + 0 562 AVAIL (16 4) (70 3)		
(16) COREQ = TELBTU − GELBTU		
(17) CELBTU = −254 1 + 0 8656 COREQ − 804 DPOL (1 8) (67 8) (12 [illegible] $+\sum_{i=0}^{m} \beta_i (PO - PC)^{b}_{-i}$	β_0 = 2.64 (0 8) β_1 = 3 87 (1 5) β_2 = 4 72 (2 2) β_3 = 5 20 (2 3) β_4 = 5 28 (2.3)	β_5 = 4 99 (2 2) β_6 = 4 32 (2 0) β_7 = 3 26 (1 9) β_8 = 1 82 (1 8)
(18) OELBTU = COREQ − CELBTU		
PRICE BLOCK:		
(19) $PFUEL = \frac{GELBTU}{TELBTU} PGAS + \frac{CELBTU}{TELBTU} \cdot PC + \frac{OELBTU}{TELBTU} \cdot PO$		
(20) PFF = PFUEL HR/10^6		
(21) $FUEL = PFF \frac{TGFF}{TG-Z3} + PNF \frac{Z5}{TG-Z3}$		
(22) PR = 2 084 + 1 0 FUEL + 0 00238(AUCC − 4.7T) − 0 5154 log $\frac{DR}{N}$ (26 0) (2 1) (47 1)		
(23) PI = 0 778 + 1 0 FUEL + 0.00366(AUCC − 4.7T) − 0 0974 log $\frac{DI}{X}$ (9 0) (3 0) (3 5)		
(24) $UCC = \left[\frac{\Delta CN}{\Delta CN + \Delta C} \left(\frac{PNK}{PFK} - 1\right)\right] UC + UC$		
(25) AUCC = UCC + 0 9AUCC$_{-1}$		

a ALL VARIABLES ARE EXPLICITLY DEFINED IN APPENDIX 1
b A 2nd DEGREE POLYNOMIAL CONSTRAINED TO ZERO AT T − M
c A 2nd DEGREE POLYNOMIAL CONSTRAINED TO ZERO AT T + 1
d A 2nd DEGREE POLYNOMIAL CONSTRAINED TO ZERO AT T + 1 AND T − M
* INDICATES THE VALUE OF THE t-STATISTICS ALSO NOTE THE SUBSCRIPT −i INDICATES THE VALUE OF A VARIABLE LAGGED i PERIODS

$\left(\frac{K_s}{N}\right)$ depends in turn on the expected cost of owning and operating the particular appliance (P_s) relative to the GNP deflator (P), and on a proxy for permanent income, which is calculated as a weighted average of lagged values of real per capita disposable income $\left(\frac{YD}{N}\right)$. Expected relative prices are assumed to be based on a weighted average of past relative prices:

$$\frac{K_s}{N} = f\left(\sum_{i=0}^{n_s} a_{is}\frac{P_s}{P_{-i}}, \sum_{i=0}^{m_s} c_{is}\frac{YD}{N_{-i}}\right) \qquad s = 1 \ldots j \text{ appliances. } (2)^*$$

Due to inadequate appliance stocks data, estimating (2)* for each appliance and then estimating (1)* was impractical; therefore,

TABLE 2 CONTINUED

$\bar{R}^2$	S. E.	D W	ESTIMATION TECHNIQUE	SAMPLE PERIOD
0.996	60 78	1.83	OLS	1951-1970
0 998	70.81	1 77	2SLS	1953-1971
0 990	0.22	0 60	2SLS	1949-1971
0.512	0.27	0 90	2SLS	1949-1971

(2)* was aggregated across appliances, leaving the capital stocks of the various appliances implicit in the per capita disposable income and relative price variables. The stock of one appliance, central air conditioners, ($K1$), was explicitly included because it is the primary determinant of peak-shaving requirements used later in the model. The average residential price of electricity deflated by the GNP-deflator $\left(\frac{PR}{P}\right)$ serves as a proxy for the cost of owning and operating an appliance.[10] The estimation form is given in equation (3)* and the results are reported in equation (1) of Table 2:

$$\frac{DR}{N} = \alpha_0 + \alpha_1 \frac{K1}{N} + \sum_{i=0}^{m} \beta_i \frac{YD}{N_{-i}} + \sum_{i=0}^{m} \Psi_i \frac{PR}{P_{-i}} + e. \qquad (3)^*$$

Industrial and commercial demands for electricity constitute the bulk of electricity requirements. It is hypothesized that there are three major determinants of demand for these users. First, by reference to production processes, the demand for electricity depends on the economy's current level of economic activity (X) measured by real GNP. Second, as the real price of electricity $\left(\frac{PI}{P}\right)$ varies, presumably production processes are adjusted over time to be more or less electricity intensive. A third factor affecting demand is the widespread adoption of central air conditioning (KCA) in commercial establishments such as office buildings and shopping malls. Stated generally, the industrial and commercial demand for electricity (DI) can be written as follows:

---

[10] The simple correlation between the price variable selected and one based on an equal weighting of the electricity price and the appliance price index was 0.98.

$$DI = f\left(X, \frac{PI}{P} \ldots, \frac{PI}{P_{-i}}, KCA\right). \qquad (4)^*$$

The results in (3) of Table 2 are based on $(5)^*$ which is cast in terms of electricity demand per dollar's worth of real GNP. This formulation mitigates problems of multicollinearity and heteroscedasticity:

$$\frac{DI}{X} = \alpha_0 + \alpha_1 \frac{KCA}{X} + \sum_{i=0}^{n} B_i \frac{PI}{P_{-i}} + e. \qquad (5)^*$$

Since the stocks of residential and commercial air conditioners ($K1$ and KCA) are primary determinants of the summer peak-shaving demands, they were treated explicitly in (1) and (3) and are endogenized in equations (2) and (4) explicitly. Estimates using a standard stock adjustment formulation indicated a unitary adjustment rate over the period of a year. As measures of the desired capital stock, distributed lags on per capita income and real GNP were used, respectively.[11]

The conversion efficiency block:

Equation (6) states the identity that total generation (TG) by both utilities and private industries equals residential (DR) and commercial and industrial demands (DI), less imports ($Z1$), plus generation losses ($Z2$) by electric utilities. Net imports are trivial and treated as exogenous. Generation losses ($Z2$) by utilities are estimated in (5) as a simple technical relationship linking generation losses to generation by utilities ($TG - Z3$). Private industrial generation ($Z3$) is about 7 percent of total generation and has not changed appreciably in recent years. Therefore, it is taken as exogenous.

Equation (7) of Table 2 posits that the share of nuclear capacity added relative to fossil and nuclear capacity additions $\left(\frac{\Delta CN}{\Delta CN + \Delta C}\right)$ is a function of relative fuel prices $\left(\frac{PNF}{PFF}\right)$ and capital costs $\left(\frac{PNK}{PFK}\right)$ in nuclear and fossil plants. These results update the findings of MacAvoy,[12] as the observations consist of an average for each power region over the five-year periods beginning in 1958 and using data for projected installations over the period 1973 through 1977. In applying (7) in a time series forecasting context, the subscripts "—7" were arbitrarily added to reflect the substantial lags in the effects of prices on new plant capacity type.[13]

---

[11] Price series for air conditioning were unavailable and the deflated price of electricity proved nonsignificant.

[12] See [9]. The procedures by MacAvoy were followed except that the

Equation (8) is a simple linking equation stating that the change in nuclear generation ($Z5 - Z5_{-1}$) depends upon the growth in generation requirements by nuclear and fossil plants ($TG - Z3 - Z4$)[14] and the share of new nuclear capacity determined in (7). Implicit in this equation is the assumption that new fossil and nuclear plants will operate at similar utilization rates.[15]

Equations (9) and (10) merely state identities useful in subsequent equations. In equation (9), it is now possible to calculate total generation requirements by fossil fuels (*TGFF*), since at previous stages total generation requirements (*TG*) and nuclear generation (*Z5*) were determined. *Z3*, industrial generation, and *Z4*, hydroelectric generation, are treated as exogenous to the model. Equation (10) states the industry's measure of fuel efficiency—the heat rate (*HR*), which measures the total btu's (*TELBTU*) required to generate a given quantity of kilowatt hours (*TGFF*). Since only *TGFF* is determinate at this point, it is not possible to calculate either the heat rate (*HR*) or the total btu's of fossil fuel required (*TELBTU*). However, after solving for the heat rate in equation (11), it is then possible to calculate *TELBTU*.

Primary attention has been given in the model to the mechanism by which the heat rate is determined because of its direct effects upon fossil-fuel requirements. Over time, a declining heat rate has been a major source of fuel credits. For example, since 1920 the heat rate has declined by 75 percent. The interruption in this pattern in the mid-1960s caused considerable speculation. While one would expect a decrease in the absolute rate of decline as the heat rate falls to the engineering "optimal" heat rate of 8800 btu/kwh,[16] engineering limitations seem unlikely to account solely for the current stability of the heat rate at 10,500 btu/kwh. Still another explanation is that decreasing fuel prices relative to capital prices lead to more fuel-intensive designs. However, this hypothesis would appear contradictory to the findings of Seitz, who observed

---

AEC forecasts of future capacity additions were used instead of his projection technique.

[13] In forecasting exercises, the relative price of nuclear to fossil capital was indexed to equal 1.0, thereby requiring an adjustment of the constant term. Also, since in forecast applications, nuclear capacity additions are well known for five years ahead, $\frac{\Delta CN}{\Delta CN + \Delta C}$ was treated as exogenous for the first five years of the forecast and endogenized via (7) thereafter.

[14] $TG - Z3 - Z4$ is total generation requirements less generation by private industry ($Z3$) and hydroelectric generation ($Z4$), leaving the balance to be met by nuclear or fossil-fuel plants. Note that hydroelectric generation is becoming of decreasing importance and is treated as exogenous to the model.

[15] For another method of how nuclear power is assumed to be phased in see [3], p. 72.

[16] See Summers [14].

the fuel-saving nature of technological change in power plants and questioned whether the fuel-saving measure of efficiency served as the overall efficiency measure in designing plants.[17] Another explanation for the lack of additional fuel efficiency gains is that peak-shaving demands are rising faster than base-load demands, causing increased relative reliance on gas turbines and older thermal equipment, both of which have higher heat rates. Still other explanations attribute the greater use of turbines and older equipment to the delay in obtaining construction site approvals. Equations (11) through (14) attempt to place these factors in perspective.

Since the heat rate of any particular unit of generation capacity is fixed for its life and technical progress is embodied, the problem of determining the average heat rate for all units at any point in time is best viewed as a vintage capital stock problem.[18] Total generation capacity (C_T) is determined as follows:

$$C_T = \int_{-\infty}^{T} \delta^{T-t} I(t)\,dt, \qquad (6)^*$$

where δ equals one minus the annual depreciation rate and $I(t)$ is the investment rate in terms of generation capacity added in year t. The average heat rate ($\overline{HR}$) is determined in equation (7)*:

$$\overline{HR}_T = \frac{\int_{-\infty}^{T} \delta^{T-t} I(t)\, f(t)\, dt}{C_T}, \qquad (7)^*$$

where $f(t)$ describes those technical and economic characteristics affecting fuel efficiency in period t, such as the process design and the relative use of different vintages of capacity. By differentiating equation (7)* with respect to T, one obtains (8)* which can be simplified as in (9)*:

$$\frac{d\overline{HR}}{dT} = \qquad (8)^*$$

$$\frac{[I(T)\, f(T) + (\ln \delta)\, C\, \overline{HR}]C - [I(T) + (\ln \delta) C]C\, \overline{HR}}{C^2}$$

$$\frac{d\overline{HR}}{dT} = \frac{I(T)}{C}\,[f(T) - \overline{HR}]. \qquad (9)^*$$

Since replacement investment in a rapidly expanding industry with a capital life of thirty years is trivial, it is assumed that

$$\frac{\Delta C}{C} = \frac{I(T)}{C}. \qquad (10)^*$$

The average heat rate ($\overline{HR}$) is measured by HR_{-1}, and $f(T)$ is a linear function of the adjusted user cost of capital relative to fuel cost $\left(\frac{UCCP}{PFUEL}\right)$ as shown in (11)*:

$$f(T) = \alpha_0 + \alpha_1 \frac{UCCP}{PFUEL}. \qquad (11)^*$$

---

[17] See [13].

[18] I owe to Warren Moskowitz of Exxon this interpretation of the problem.

The user cost of capital equipment for utilities (UC) is forecast by the Wharton Annual Model and serves as a basis for adjustment.[19] In using capital costs relative to fuel costs as an explanatory variable, it is frequently overlooked that changing rates of capacity utilization affect the relative prices. For example, higher rates of capital utilization imply lower capital cost per unit of output and thereby result in a lower heat rate as new installations favor fuel-saving process designs in response to the higher relative price of fuel. In recent years, the advent of air conditioning has produced a situation in which peak-period demands are growing at a faster rate than base-load demands, causing the average utilization of capacity to fall. The user cost of capital is adjusted in equation (14) to reflect this phenomenon.[20]

Assuming for the moment that the ratio of peak to base-load demands $\left(\frac{Q_p}{Q_b}\right)$ is known, equations (10)*, (11)*, and (14) can be substituted into (9)*, giving the following estimation form:

$$HR - \frac{C_{-1}}{C} HR_{-1} = \alpha_0 \frac{\Delta C}{C} + \alpha_1 \frac{UCCP}{PFUEL} \frac{\Delta C}{C} + e. \qquad (12)^*$$

Both $UCCP$ and $PFUEL$ were treated as endogenous variables. The 2SLS regression results are given in equation (11) of Table 2.

In order to calculate $UCCP$ in (14), it is necessary to determine the ratio of the average monthly output during the summer peaking season to the average monthly output for the year $\left(\frac{Q_p}{Q_b}\right)$. It is hypothesized that the summer peak demand for air conditioning is a direct function of the stock of residential and commercial air conditioners ($K1 + KCA$) as follows:

$$\left(\frac{Q_p}{Q_b} - 1\right)\Psi(TG - Z3) = \alpha_0 + \alpha_1 (K1 + KCA) + e. \qquad (13)^*$$

$TG - Z3$ is the annual generation by the electric utility industry as it excludes industrial generation ($Z3$). The parameter Ψ, which is included to show that the peak load demands only occur during a fixed portion of the year, becomes implicit in the estimated right-hand-side parameters given in equation (13) of Table 2.

The fossil-fuel block:

The estimation of the demand for natural gas by electric utilities is confounded by the fact that interstate prices probably are not

---

[19] The adjustment process is given in equation (14).

[20] Equation (14) is based on the following:

$$UCCP = UC\left[1. + \frac{U_b}{U_p}\left(\frac{Q_p}{Q_b} - 1.\right)\right].$$

U_b and U_p are the annual utilization rates of base and peak load units, respectively, and Q_p and Q_b are the average monthly output during the summer peaking season and the average monthly output for the year, respectively. $\frac{U_b}{U_p}$ was assumed to equal 6 because the summer peaking period is essentially of two months' duration.

market clearing.[21] Consequently, purchases by utilities may be determined by supply considerations which in turn depend on the demand by other types of natural gas users. Due to such problems, a variety of polynomial lags using natural gas, coal, and fuel oil prices provided no evidence for a price response. Consequently, a linking equation was developed that implies that the natural gas sold to electric utilities (*GELBTU*) depends on supply limitations —the availability of "low priority" gas (*AVAIL*). The availability of "low priority" gas[22] includes *GELBTU*, refinery uses, and transmission and field uses. Equation (15) explicitly introduces supply constraints and a fixed marginal relationship between *GELBTU* and the availability of low priority gas. In future forecasting exercises, *AVAIL* serves as an approximation to the variable the FPC will likely choose to ration most stringently.

Coal, still the dominant fuel for electricity generation, has faced increasing competition from residual fuel oil along the eastern seaboard and from natural gas in the inland areas. For inland markets, the extent of competition between coal and oil is much more limited depending primarily on nearby refining sources. As a result, in interpreting aggregate price responses for the United States as a whole, one must recognize that these data reflect both a weighting across various local and geographically distinct markets and a weighting of "spot" and long-term contract prices.

As indicated in equation (17), three factors determine the demand for coal. First, the extent of coal's potential market (*COREQ*) is determined by subtracting the sales of natural gas (*GELBTU*) from total btu requirements (*TELBTU*) in equation (16). Second, the lagged values of the price differential between fuel oil and coal ($PO - PC$) describe interfuel substitution possibilities. Finally, a dummy variable for sulphur emission controls (*DPOL*) beginning in 1969 has been introduced.

Since the demand for coal is determined from the total amount of fuel to be met by coal and/or residual fuel oil (*COREQ*), residual fuel demand is necessarily determined in equation (18). This is reasonable, both mathematically and logically as residual fuel oil is generally the marginal source of supply as witnessed by the rapid growth of fuel oil demand in response to shortages of natural gas and low sulphur coal. The treatment of residual fuel oil demand as a mathematical residual does not imply that it is exogenous to the system; to the contrary, lower fuel oil prices in (17), *ceteris paribus*, will increase fuel oil demand by reducing coal demand (*CELBTU*).

The price block:

Before proceeding to estimate the prices of residential and commercial and industrial electricity service, it is necessary to calculate fuel costs on a kilowatt hour basis. Equation (19) calculates an average price for fossil fuel on a btu input basis by weighting the respective exogenous fuel prices by their relative shares. Equation (20) derives *PFF*, the price of fossil fuel on a kilowatt hour basis by applying the heat rate (*HR*) to the price in btu's (*PFUEL*). In equation (21), the price of all fuel per kilowatt hour (*FUEL*) is found by weighting the fossil-fuel price (*PFF*) and nuclear fuel (*PNF*) by their shares of generation.

---

[21] See Wellisz [17] for a description of the procedure for rate determination.

[22] The term "low priority" gas should not be confused with interruptible sales, as most gas sold to utilities is noninterruptible. "Low priority" is used to indicate that its use is for boiler fuel, transmission fuel, etc. If the FPC assumes control over intrastate sales, these uses are likely to be termed "low priority" and become a prime target for rationing.

The average price for residential electricity service is hypothesized to depend upon a markup over unit fuel costs (*FUEL*), unit labor costs (*ULC*), capital costs (*AUCC*), and the quantity purchased per customer. Since the number of residential customers is not endogenous to the model, residential demand per capita $\left(\frac{DR}{N}\right)$ was used as a proxy as shown below:[23]

$$PR = \alpha_0 + \alpha_1\, FUEL + \alpha_2\, ULC + \alpha_3\, AUCC + \alpha_4 \log \frac{DR}{N} + e. \qquad (14)^*$$

The results are reported in equation (22).[24] In determining rate schedules, regulatory agencies base their calculation on historical capital costs. After adjusting the user cost of capital for the price differential between nuclear and fossil-fuel capital in (24), a weighted average of past values of the adjusted user cost of capital is calculated in (25), which measures historical capital costs. The weighted capital cost variable was then detrended by estimating this series as a function of a time trend, which serves as a rough measure of technological change.

The average price of commercial and industrial electricity is based on the same formulation as in (14)*, except demand per dollar of real GNP $\left(\frac{DI}{X}\right)$ serves as a proxy for increasing customer demand. The results are shown in equation (23).

☐ **Comparison of elasticity estimates.** One test of a model is the plausibility of parameter estimates in the individual equations. Table 3 reports elasticities calculated about sample means for the short run and long run, where the long run is defined as the full impact after all lagged effects are fully realized. These results uniformly suggest that the short-run response to changes in either factor input prices or electricity prices is low and that the long-run effects are substantial.

A number of studies of residential electricity demand have been performed using either cross-section or pooled data. Some of the more recent studies include those by Wilson, Mount *et al.*, by

---

[23] The logarithm of per capita demand was used to provide a near constant partial elasticity, as this appeared more compatible with the tariff schedules given in *Typical Electric Bills*.

[24] In view of regulatory practices based on allowing full rate adjustment to cover fuel costs, α_1, was taken to equal unity, *a priori*. Unit labor costs (*ULC*) for utilities from the Wharton Long-Term Model proved to be nonsignificant, as they changed very little over the period.

TABLE 3

MEAN ELASTICITY ESTIMATES

VARIABLES	IMPACT ELASTICITY (1 YEAR)	LONG-RUN ELASTICITY
DR| PR/P	−0.06	−0.52
DR| YD/N	0.06	0.88
DR| K1	0.22	0.22
K1| YD	0.25	6.46
DI| PI/P	−0.04	−0.51
DI|X	0.89	0.89
DI| KCA	0.13	0.13
KCA| X	0.34	3.16
$\frac{\Delta CN}{\Delta CN + \Delta C} \mid \frac{PNF}{PFF}$	0	−1.40
$\frac{\Delta CN}{\Delta CN + \Delta C} \mid \frac{PNK}{PFK}$	0	−0.86
$HR \mid \frac{\Delta C}{C}$	0.15	1.45
$HR \mid \frac{UCCP^a}{PFUEL}$	0.03	0.30
GELBTU| AVAIL	1.33	1.33
CELBTU| CORE Q	0.99	0.99
CELBTU| PC	−0.02	−0.20
PR| AUCC	0.15	0.15
$PR \mid \frac{DR}{N}$	−0.21	−0.21
PI| AUCC	0.45	0.45
$PI \mid \frac{DI}{X}$	−0.08	−0.08

[a] ASSUMES $\frac{\Delta C}{C} = 0.10$.

Anderson, and by Houthakker *et al.* With the exception of Fisher and Kaysen's findings,[25] price elasticity estimates range from −0.9 to −1.3, suggesting that the estimate obtained here of a −0.52 price elasticity understates the true impact.[26]

The elasticity estimates in this study are derived differently from those referred to above in the three following respects:[27] (1) the use of pure time series observations, (2) the use of polynomial distributed lags, and (3) the use of the average price of electricity rather than approximate marginal tariff rates. The usual objections to using average prices involve problems of simultaneous equation bias[28] and the possibility that changes in average rates may not serve as a good proxy for changes in marginal rates. This study, which utilizes two-stage least squares, overcomes the first objection, while the second seems improbable given the extent of aggregation. In addition, the author feels that the polynomial distributed lag approach offers a distinctly more plausible adjustment pattern than geometric lag formulations, which place primary weight on the current period. However, it is true that time series estimates (even with distributed lags) generally tend to understate the price response so the estimates here may form a lower bound. In addition, Chow tests for stability, which were run for the major equations in the model, revealed a possible instability in the estimators for this equation.

---

[25] See [7].

[26] Wilson's [18] estimate is −1.33. Anderson [2] finds a price elasticity of −0.91, Mount, Chapman and Tyrrell's estimate [10] is −1.20, and Houthakker, Verleger, and Sheehan [8] get −1.0.

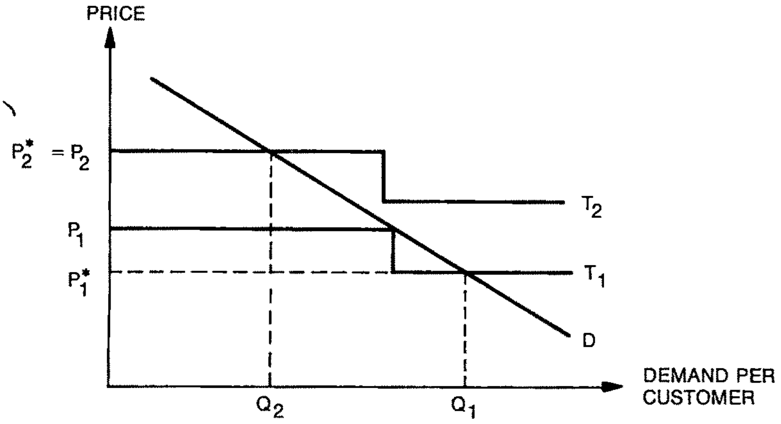
[27] Wilson found a substantial cross elasticity of electricity demand, while Anderson's estimate is 0.13 and Mount, Chapman, and Tyrrell get 0.19. This study, like that of Houthakker *et al.*, found no significant effect of natural gas prices.

[28] Basically, an increase in demand per consumer moves him to a lower marginal tariff rate, which causes the average price to fall. See Anderson [2] for a full discussion.

Studies using marginal tariff rates will typically result in overestimates of the true price response, so that the true price elasticity probably lies somewhere between —0.5 and —1.0. Consider Figure 2, which depicts a downward sloping demand curve D and two tariff schedules (T_1, T_2) for different periods of time. If one uses the marginal rates for a fixed output as in the above studies, the elasticity of demand is calculated as $n = \frac{Q_2 - Q_1}{Q_2} \frac{P^*_2}{P^*_2 - P_1}$. The proper point elasticity estimate is based on the *true* price differential as follows: $n^* = \frac{Q_2 - Q_1}{Q_2} \frac{P^*_2}{P^*_2 - P^*_1}$. Since $P^*_1 < P_1$, $|n^*| < |n|$, suggesting that the studies using tariff schedules are biased and probably form an upper bound on the price response.

FIGURE 2

PROBLEMS WITH MARGINAL TARIFF SCHEDULES



Unfortunately, there are only a few sources against which to compare the remaining estimates. At first glance, the industrial and commercial price elasticity estimates might be expected to be absolutely greater than the residential estimate in view of the alleged tendency towards price discrimination against the residential user. It should be pointed out that the industrial elasticity for electricity generated by utilities may be much higher as industrial plants may construct their own generation facilities, but this would only result in the substitution among sources of generation and need not imply greater aggregate price responsiveness. Elasticity estimates by Mount, Chapman, and Tyrrell for commercial and industrial demands range from —1.36 to —1.82 for their variable elasticity model calculated about sample means.[29] Again this suggests the possibility of the understatement of price effects, but for the above reasons, this is by no means certain. The variables reflecting increased usage per customer on *PR* and *PI* compare favorably with calculated values of —0.23 and —0.13, respectively, based on tariff schedules.[30]

---

[29] See [10].

☐ **Sample period simulations.** While favorable results from sample period simulations are not sufficient evidence of a model's forecasting ability, they are probably a necessary condition and are instructive in testing the various equations' behavior in the overall system. Over the period 1953 to 1971, the model is simulated using actual exogenous variables and calculated lagged endogenous variables for years after 1953. These simulations are based on the complete model given in Table 2 with two exceptions. First, because of the modest amount of nuclear generation over the period, it is exogenized. Second, actual values of the rate of new fossil capacity additions are used in place of the scheme given in (12), since in forecast applications capacity additions can be anticipated several years in advance.

In Table 4, average absolute percentage errors for selected variables are reported for both the full sample period simulation and a "trend model." The "trend model" is based upon a logarithmic trend line extrapolation using the actual values for the beginning year, 1953, and assumes that values for the ending year, 1971, were somehow known in 1953. The econometric model clearly out-performs the naive trend-line forecasts. However, for residential demand (*DR*) and total generation (*TG*), the differences are not great; the so-called 7-percent growth rule for predicting total generation performed well over the period. The conversion efficiency and fossil-fuel blocks add much greater realism than the trend-line procedure. The pattern of declining heat rates until the mid-1960s and, since then, of moderately rising rates, is not well-approximated through the trend-line method. The forecast error for *TELBTU* is almost two-thirds larger than the simulation model. Similarly, changing relative fuel prices and pollution regulations produced substantial changes in the mix of fuels burned which are not well explained by the trend model. Both models have difficulty

TABLE 4

COMPARISON OF AVERAGE ABSOLUTE PERCENTAGE ERRORS FOR FULL SAMPLE PERIOD (1953-1971)

	THE MODEL	NAIVE TREND MODEL
RESIDENTIAL DEMAND (DR)	2.2	3.0
INDUSTRIAL DEMAND (DI)	2.1	3.2
TOTAL GENERATION (TG)	1.6	2.4
TOTAL FOSSIL FUEL REQUIREMENTS (TELBTU)	2.4	3.9
COAL CONSUMED (CELBTU)	3.7	7.0
RESIDUAL FUEL CONSUMED (OELBTU)	6.7	64.4

---

[30] Using *1971 Typical Electric Bills*, the arc elasticity for 500 and 1000 kwh per month is —0.23. An arc elasticity calculation for 30,000 and 60,000 kwh is —0.13.

forecasting fuel oil demand. The large errors can be attributed both to the role of residual fuel oil as a mathematical residual and to its small share of total fuel consumed.

3. Policy simulation exercise

■ In the introduction, the question was posed how changes in relative fuel prices and generally higher fuel prices will affect the mix of fuels consumed, the price of electricity, and the demand for electricity. In order to consider this question it is necessary to prepare a base case forecast as a means of comparison with simulations of differing fuel prices.

In forecasting applications, all of the equations given in Table 2 are utilized. Equation (7), the share of nuclear capacity, is exogenized until 1978, after which nuclear additions are not known with much certainty and (7) becomes operative. The values for all exogenous variables are given in detail in Appendix 2. (The sources of these variables include the Wharton Long-Term Model for macroeconomic variables,[31] various public and private forecasts,[32] and *ad hoc* trend projections.[33]) The model is simulated over the period 1974 through 1981, as the Wharton forecast for the macroeconomic variables extends to 1981. The macroeconomic setting over the period hypothesizes slower growth in real disposable income (3.4 percent versus 4.6 percent over the 1960s), 3.9-percent average annual growth in real GNP, 2.9-percent average annual inflation. Even in constant dollars, fuel prices are assumed to escalate at the following annual rates: nuclear 0.0 percent; residual fuel oil 3 percent; natural gas 2 percent; and coal 2 percent. As indicated in Appendix 2, these rates of increase are quite substantial in current dollars. These assumptions provide the exogenous inputs for Case A, our base case simulation. Given in Table 5 are the results for both Case A and the alternate Case B.

Before comparing the two cases, let us consider the forecast implications of Case A. Note in particular that residential demand is predicted to grow at a considerably slower rate than in the past

---

[31] The macroeconomic variables are taken from the December 1972, "Oil Imports Adjustment, Medium World Trade" solution to the Wharton Long-Term and Industry Model of the U. S. Economy. These variables include real disposable personal income (*YD*), real GNP (*X*), GNP deflator (*P*), and user cost of capital in utilities (*UC*).

[32] Population (*N*) was based on an average of Series A and B in *People of the U. S. in the 20th Century* (Census Monograph, Washington, D. C., 1971) by Taeuber and Taeuber. "Low priority" natural gas (*AVAIL*) is based on a forecast by a private petroleum firm. The relative price of nuclear capital to fossil capital seems consistent with a January 1973 article entitled "Capital Cost Calculations for Future Power Plants" in *Power Engineering*. The prices of nuclear fuel (*PNF*), fuel oil, coal, and natural gas are all based on the National Petroleum Council's 1973 Report. Case B uses an average of cases 1 and 2 for all fuel prices. Case A assumes half the real price increases of Case B.

[33] The sulphur pollution control dummy is set to current levels (2.0) for the remainder of the period. Note that a relaxation of standards from 2.0 to 1.0 would imply a reduction of oil demand by $804 \cdot 10^{12}$ btu's or approximately 25 percent. In addition, this change might be implemented in the short run and would imply an increase in coal demand of about 9 percent. Net imports (*Z*1), industrial generation (*Z*3), and hydroelectric generation (*Z*4) are based on trend extrapolation.

TABLE 5

RESULTS OF FORECAST SIMULATIONS (1974-1981)

VARIABLE	CASE	1974	1975	1976	1977	1978	1979	1980	1981	AVERAGE % GROWTH, 1974-1981
RESIDENTIAL DEMAND (DR)	A	577	614	646	676	704	733	765	800	4.8
	B	577	613	645	674	703	732	763	797	4.8
INDUSTRIAL DEMAND (DI)	A	1275	1325	1404	1482	1544	1611	1693	1786	4.9
	B	1272	1320	1396	1472	1530	1594	1672	1762	4.8
TOTAL GENERATION (TG)	A	2020	2114	2234	2351	2448	2553	2674	2813	4.8
	B	2016	2108	2224	2337	2431	2531	2649	2783	4.8
NUCLEAR GENERATION (Z5)	A	101	119	145	176	211	266	330	410	22.1
	B	100	118	143	173	207	261	327	410	22.2
HEAT RATE (HR)	A	10523	10501	10465	10429	10406	10387	10365	10338	−0.2
	B	10503	10469	10418	10369	10337	10311	10282	10250	−0.3
NATURAL GAS (GELBTU)	A	3678	3565	3453	3341	3229	3116	3003	2891	−3.5
	B	3678	3565	3453	3341	3229	3116	3003	2891	−3.5
COAL (CELBTU)	A	9298	10051	10950	11770	12381	12883	13407	13978	6.0
	B	9268	10005	10884	11685	12303	12818	13345	13900	5.9
FUEL OIL (OELBTU)	A	3011	3029	3063	3081	3068	3048	3046	3046	0.1
	B	2987	2981	2967	2953	2889	2801	2725	2643	−1.2
RESIDENTIAL PRICE (PR)	A	2.16	2.17	2.17	2.18	2.19	2.21	2.21	2.22	0.4
	B	2.20	2.21	2.23	2.25	2.27	2.29	2.31	2.33	0.8
INDUSTRIAL PRICE (PI)	A	1.41	1.44	1.47	1.50	1.52	1.55	1.57	1.59	1.7
	B	1.45	1.50	1.53	1.57	1.60	1.64	1.67	1.70	2.3

because of slower macroeconomic growth and the declining elasticity of air conditioning with respect to disposable income.[34] Nuclear generation is predicted to grow at a 22.1-percent annual rate, and by 1981, almost 15 percent of generation requirements would be met by nuclear power. Gains in fossil-fuel efficiency should reappear, but proceed at only a modest rate, as the growth in fossil-fuel prices tends to be largely offset by lower capacity utilization due to increasing peak-shaving demands. Among fossil fuels, the supply limitations on natural gas and the marked increase in fuel oil prices imply that coal demand will grow significantly over the period despite the rapid growth in nuclear generation. The prices of electricity, which have historically declined, will increase only modestly for residential service and at somewhat faster rates for commercial and industrial service.

In order to assess the impact of higher fuel prices and greater price differentials among fuel inputs, an alternative solution (Case B) to the model was calculated assuming higher prices for fossil fuels, which are more consistent with NPC projections. Case B assumes the following annual rates of constant dollar escalation in fossil-fuel prices: residual fuel oil 4.6 percent; natural gas 3.9 percent; and coal 3.9 percent. As Appendix 2 indicates, the differences between Cases A and B are substantial, especially in the late 1970s.

---

[34] The income elasticity is 5.7 based on 1965 values, and declines to 2.3 by 1980. A similar but less dramatic change occurs for commercial air conditioning with respect to real GNP as it declines from 3.1 in 1965 to 1.6 in 1980.

In comparing Cases A and B, two important results are indicated.[35] First, higher fossil-fuel prices in Case B have a relatively small impact on electricity prices and an even smaller effect upon electricity demand. Second, the effects on the mix of fuel inputs are significant and reveal interesting interrelationships not apparent *a priori.*

The price of residential electricity is only 5 percent higher in 1981 in Case B compared to Case A despite the fact that oil, coal, and natural gas prices increased by 30 percent, 19 percent, and 21 percent, respectively. The effect on residential electricity demand is negligible due in part to the declining price elasticity over the extended sample range. For industrial and commercial users, the effects of higher fuel prices are relatively greater but still rather small. In 1981, the average industrial and commercial electricity price is 7 percent higher and demand is 1.3 percent lower as the effects of higher fuel prices are recognized only after a substantial lag.

The increasing price of fossil fuels relative to nuclear fuels results in an increase in the mix of nuclear generation. The decreased total generation requirements resulting from lower electricity demands offset the nuclear mix augmenting effect, leaving nuclear generation essentially unchanged in the two cases. With a shift in generation mix to favor nuclear, one would certainly expect lower demands for residual fuel oils and possibly lower coal demands. As indicated by the 1-percent lower heat rate in 1981, resulting from higher fossil-fuel prices, fewer fossil-fuel inputs will be needed to achieve a stated output. In 1981, coal demands are negligibly lower (0.6 percent) in Case B but fuel oil demands are reduced by 13 percent. Moreover, the impact will no doubt be much greater after the effects of the increasing differential between coal and fuel oil prices are fully felt in subsequent periods.

Thus the above results suggest that when the markedly higher fossil fuel prices are translated into increases in electricity prices, the effects are likely to be small and will have an impact on demand only after a substantial period. It therefore seems unlikely that somewhat higher fuel prices will alleviate short-term shortage situations through reducing final consumption of electricity. The effects of changing relative fuel prices can, however, be expected to induce substitution of other lower priced fuels for those in short-supply, although this too is not an instantaneous process.

The model would appear more limited for forecasting than for simulation use for several reasons. Structural shifts and linear approximations to nonlinear phenomena over a restricted sample range pose potentially serious problems, especially in applications over longer time horizons. Another problem is the uncertainty attached to the exogenous variables. This is particularly true concerning factors affecting fuel inputs choices. Environmental concerns over sulphur emissions, nuclear licensing delays, and political considerations affecting the price and availability of petroleum products, along with the usual technological uncertainties, compound the difficulties. Simulation exercises are by no means immune to such problems, but appear much less sensitive than forecasting applications.

---

[35] Since these simulations were performed, fossil-fuel prices have risen much more sharply than anticipated, making the contrast between fuel prices in Cases A and B appear only modest in retrospect. Obviously, these higher realized fuel prices will imply greater reductions in electricity consumption. Nevertheless, the essential point of these simulations—that the elasticity of electricity demand with respect to fuel input prices appears highly inelastic, especially in the short run—remains valid.

The above exercise contrasting the effects of higher fossil-fuel prices is only one possible application of the model for policy purposes. For example, by adjusting the user cost of capital, the model could consider the effects of higher capital costs, tighter zoning restrictions, and plant redesign expenditures. Still another question is the effect of a sulphur tax on the fuel inputs and demand for electricity.[36] Thus, the model seems sufficiently detailed so as to allow simulation studies on a number of policy issues.

Appendix 1

■ The variables used in the model and the sources of the data are listed below.

$AUCC$ = Weighted average of user cost of capital for utilities.

$AVAIL$ = The difference between total demand and residential, commercial and industrial uses, where industrial uses exclude refinery fuel, field uses, and transmission uses but include carbon black use (in 10^{6} mcf's). (U. S. Bureau of Mines, *Statistical Yearbook.*)

C = Conventional steam plus internal combustion generation capacity. (Edison Electric, *Statistical Yearbook*, hereafter E.E., Table 2S.)

$CELBTU$ = 10^{12} btu's of coal consumed by the electric utility industry (calculated by using btu/ton data from Edison Electric and Federal Power Commission data on tons consumed).

CN = Nuclear generating capacity. (E.E., Table 2S)

$COREQ$ = Coal and oil requirements in 10^{12} btu's.

DI = Industrial demand includes both sales of small and large light and power, street and highway lighting, other public authorities, railroads, interdepartmental plus total self-generation in 10^{9} kwh. (E.E., Table 7.1S and 19S.)

$DPOL$ = Dummy sulphur pollution control variable ($1969 = 0.5$; $1970 = 1.0$; $1971 = 1.5$).

DR = Residential sales by the electric utility industry in 10^{9} kwh. (E.E., Table 19S.)

$FUEL$ = Average fuel costs of generating a kwh of electricity (E.E., Table 53S.)

$GELBTU$ = 10^{12} btu's of natural gas purchased by the electric utility industry calculated by using the Edison Electric's average annual conversion factor. (E.E., Table 42S.)

---

[36] The effects of a sulphur tax are explored in a forthcoming paper in the *Journal of Political Economy.*

HR = Heat rate or the btu's required to generate a kilowatt hour. Note that this series differs from that estimated by the Federal Power Commission. Instead of using constant coefficients to convert the fuel's volumes or weights to btu's, the weights were allowed to vary based on the conversion factors in E.E., Table 42S. This allows for the use of lower viscosity, low-sulphur fuel oil and Western, low btu, low-sulphur coal.

$K1$ = Stock of residential central air conditioners in 10^3 tons. (See Table 6 of Statistical Release Mimeo of April 19, 1971, by Air Conditioning and Refrigeration Institute.) The stock was calculated from residential shipments data assuming a life of 8 years; for years prior to 1956, total shipments given in Table 1 were used to approximate the series. Note that window unit data are not included as they were unavailable.

KCA = Stock of air conditioners owned by commercial and industrial users in 10^3 tons. See $K1$ for details.

N = Population in millions.

$OELBTU$ = Fuel oil consumed by utilities in 10^{12} btu's. (E.E., Table 42S.)

P = GNP deflator, 1958 = 1.0.

PC = Average cost of coal to utilities in ¢/10^6 btu's. (E.E., Table 42S.)

PFF = Cost per kwh of fossil fuel (¢/kwh).

PFK = Cost per megawatt of new fossil capacity.

$PFUEL$ = Price of fuel in ¢/10^6 btu's. (E.E., Table 42S.)

$PGAS$ = Average cost of natural gas to utilities in ¢/10^6 btu's. (E.E., Table 42S.)

PI = A weighted average price for service to small and large light and power users in ¢/kwh. (E.E., Table 45S.)

PNF = Cost per kwh of nuclear fuel (¢/kwh).

PNK = Cost per megawatt of new nuclear capacity.

PO = Average cost of natural gas to utilities in ¢/10^6 btu's. (E.E., Table 42S.)

PR = Average revenue per kilowatt hour sold to residences in ¢/kwh. (E.E., Table 45S.)

Q_b = Annual average monthly output in 10^9 kwh. (E.E., Table 16S.)

Q_p = Average monthly output for July and August in 10^9 kwh. (E.E., Table 16S.)

T = Time trend 1942 = 1.0.

$TELBTU$ = Total fossil-fuel requirements in 10^{12} btu.

TG = Total generation by electric utilities and industrial generators in 10^9 kwh. (E.E., Table 7.)

$TGFF$ = Total generation in the electric utility industry with fossil fuels in 10^9 kwh. (E.E., Table 41S.)

UC = User cost of capital equipment for utilities, e.g., 1970 = 31.7 (Wharton [12].)
UCC = User cost of capital equipment adjusted for differential prices of nuclear and fossil capital.
$UCCP$ = User cost of capital adjusted for peaking uses.
X = GNP in 10^9 1958 \$.
YD = Disposable income in 10^9 1958 \$.
$Z1$ = Net imports in 10^9 kwh. (E.E., Table 7.)
$Z2$ = Transmission losses in 10^9 kwh. (E.E., Table 7.)
$Z3$ = Generation by industrial plants in 10^9 kwh. (E.E., Table 7.)
$Z4$ = Hydroelectric generation in 10^9 kwh. (E.E., Table 13S.)
$Z5$ = Nuclear generation in 10^9 kwh. (E.E., Table 13S.)

Appendix 2

■ Table A1 shows the exogenous variables for forecast simulation in Case A and Case B.

TABLE A1

EXOGENOUS VALUES FOR FORECAST SIMULATION

CASE A

	N	YD	P	X	PNF	Z4
1972	211.7	575.4	1.459	789.2	0.176	275.3
1973	216.4	612.6	1.503	838.1	0.181	284.7
1974	221.1	640.8	1.553	868.9	0.186	294.4
1975	225.8	647.4	1.623	890.0	0.191	304.4
1976	230.0	668.6	1.660	933.9	0.196	314.7
1977	234.2	702.3	1.720	977.1	0.203	325.4
1978	238.5	722.0	1.786	1005.4	0.208	336.5
1979	242.7	745.4	1.857	1038.2	0.214	347.9
1980	246.9	772.2	1.905	1081.4	0.221	361.8
1981	251.5	806.9	1.961	1133.3	0.227	372.0
	UC	AVAIL	PO	PC	PG	T
1972	28.3	8200	58.7	38.1	30.5	31
1973	29.1	8000	62.2	39.9	32.0	32
1974	30.4	7800	65.8	41.9	33.6	33
1975	32.8	7600	69.7	43.9	35.2	34
1976	32.3	7400	73.9	46.1	36.9	35
1977	32.5	7200	78.2	48.3	38.8	36
1978	33.6	7000	82.8	50.7	40.6	37
1979	34.6	6800	87.7	53.2	42.6	38
1980	35.0	6600	92.8	55.8	44.7	39
1981	35.5	6400	98.3	58.5	46.9	40

Z1 = 3.0
PNK/PFK = 1.0
Z3 = 105
DPOL = 2.0

CASE B

	PO	PC	PG
1972	60.4	38.8	31.1
1973	65.7	41.5	33.3
1974	71.5	44.3	35.5
1975	77.8	47.7	37.9
1976	84.7	50.7	40.6
1977	92.4	54.2	43.4
1978	100.6	57.9	46.3
1979	109.5	61.9	49.5
1980	119.3	66.2	53.0
1981	127.8	70.7	56.6

References

1. Almon, S. "The Distributed Lag Between Capital Appropriations and Expenditures." *Econometrica*, Vol. 33 (January 1965), pp. 178–196.
2. Anderson, K. "Residential Demand for Electricity: Econometric Estimates for California and the U. S." Rand Corporation, R-905-NSF, January, 1972.
3. Data Resources, Inc. "The Energy Balance Model: A Preliminary Report." Lexington, Mass.: Data Resources, Inc., November 1972.
4. Dhrymes, P. *Econometrics.* New York: Harper & Row, 1970.
5. Duchesneau, T. D. "Interfuel Substitutability in the Electric Utility Sector of the U. S. Economy." *Economic Report to the Federal Trade Commission.* February 1972.
6. Edison Electric Institute. *Statistical Yearbook of the Electric Utility Industry.* New York: Edison Electric Institute.
7. Fisher, F. M. and Kaysen, C. *A Study in Econometrics: The Demand for Electricity in the United States.* Amsterdam: North-Holland, 1962.
8. Houthakker, H., Verleger, P., and Sheehan, D. "Dynamic Demand Analyses for Gasoline and Residential Electricity." Paper presented at December 1973 A.E.A. Meetings.
9. MacAvoy, P. W. *Economic Strategy for Developing Nuclear Breeder Reactors.* Cambridge: M.I.T. Press, 1969.
10. Mount, T. D, Chapman, D., and Tyrrell, T., J. "Electricity Demand in the United States: An Econometric Analysis." Papers and Proceedings of the February 1973 Conference at M.I.T. on Energy, Demand, Conservation, and Institutional Problems.
11. National Petroleum Council. *United States Energy Outlook.* Washington, D. C.: December 1972.
12. Preston, R. *The Wharton Long-Term and Industry Model of the U. S. Economy.* Philadelphia: Economics Research Unit, University of Pennsylvania, 1973.
13. Seitz, W. D. "Productive Efficiency in the Steam Electric Generating Industry." *Journal of Political Economy*, Vol. 79, No. 4 (July/August 1971), pp. 878–886.
14. Summers, C. M. "The Conversion of Energy." *Scientific American* (September 1971), pp. 149–160.
15. U. S. Federal Power Commission. *Typical Electric Bills.* Washington, D. C.: U. S. Government Printing Office, 1971.
16. Verleger, P. "An Econometric Analysis of the Relationships between Macro Economic Activity and U. S. Energy Consumption," in Milton Searl, ed., *Energy Modeling*, Washington, D. C.: Resources for the Future, 1973.
17. Wellisz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy*, Vol. 71, No. 1 (February 1963), pp. 30–43.
18. Wilson, J. W. "Residential Demand for Electricity." *Quarterly Review of Economics and Business*, Vol. 2, No. 1 (Spring 1971), pp. 7–22.

An economic model of the world oil market

Michael Kennedy

Assistant Professor of Economics and Middle Eastern Studies
University of Texas at Austin

This paper presents a regional multicommodity economic model of the world oil market. The four sectors of the model are crude production, transportation, refining, and consumption of products. The regions studied are the United States, Canada, Latin America, Europe, the Middle East and Africa, and Asia. In the model the exogenous variables include the regional supply and demand equations, the technology of refining, and government policy variables. The scope of these variables gives the model sufficient information to determine the levels of consumption, production, and price for each commodity in each region, the pattern of world trade flows, and the refinery capital structure and output in each region. The consequences of monopolistic behavior—implemented through the producing country cartel, OPEC—can be studied through the simulation of the effects of changes in the export tax. Use of this model shows that the price increases for crude oil which occurred in late 1973 are not likely to persist because the largest producing region—the Persian Gulf and North Africa—would have the problem of allocating reduced production levels among the individual countries.

1. Introduction

■ The course of world events in late 1973 has made it unnecessary to explain the need for a working model of the world oil market. Projections of consumption, pricing, and trade patterns of oil in future years are important inputs into business and government assessment of the likely economic situation. This paper describes the structure of an economic model of the world oil market, and presents some results from it for use in forecasting and policy simulation.[1]

Michael Kennedy received the A.B. from the University of Michigan (1970) and the Ph.D. from Harvard University (1974). His research is in the area of applied microeconomics.

This paper is taken from the author's Ph.D. dissertation in economics at Harvard University. I am deeply grateful to my principal adviser, Hendrik S. Houthakker. He first suggested the topic, then made fundamental contributions to its progress by guiding me away from potential quagmires and toward promising areas. Paul MacAvoy read the first version of the paper and made many useful comments and suggestions I have also received helpful advice and assistance from Joseph C. Dickson, Otto Eckstein, Timothy Greening, Dale W. Jorgenson, James Moose, and Amir Sepahban. The Ford Foundation's Energy Policy Project provided early financial support for the research. Unfortunately, none of the above can be held responsible for the errors and omissions which remain.

[1] Several versions of a world oil model are currently under development, differing by regional complexity, product breakdown complexity, and target date. Only one version is reported here.

After describing the model, indicating its relation to existing literature, and giving a mathematical presentation of the model, we shall discuss each of the sectors of the model and shall describe the results of the model's simulations.

The model which is reviewed here is a regional multicommodity economic model. It is an economic model because both the prices and the quantities transacted of oil and refined products are endogenous variables. The exogenous variables are the factors governing the conditions of supply and demand in the oil market. Some of these factors are world income growth, the costs of construction of refinery and transportation equipment, the availability of substitute fuels as it affects price elasticity, and government policies such as import or export duties. Thus, the model is a partial equilibrium model in which oil prices do not affect other economic quantities. The time framework of the model is annual, so that the prices and quantities transacted generated in any solution refer to a given year.

The version of the model reviewed here is static and long-run in nature. Since it is static, equilibrium values of the endogenous variables depend only on the levels of the exogenous variables in the solution year, and not on the past values of any variables. Thus, the model can generate the final equilibrium resulting from change in exogenous conditions, but not the time path of variables as the equilibrium is approached. The long-run orientation means that the implicit behavior of participants in the oil market, as embodied in the conditions of supply and demand, reflects a total adjustment to the levels of exogenous variables and prices prevailing in that year. This, of course, implies that capital stocks can be fully adjusted to desired levels implied by long-run optimizing behavior on the part of economic agents. These considerations make the model inappropriate for short time horizons, when many capital stocks are effectively fixed and when behavior is partially a reflection of previous conditions in the market. Instead, the model is used for intermediate range projections of five to fifteen years, over which reactions to changes in exogenous variables can be fully worked out.

The model consists of four segments: crude production, transportation, refining, and consumption of products. Commodities are distinguished both by physical characteristics (with crude oil and the various refined products explicitly represented) and by location. In each region the demand for refined products and the supply of crude oil are functions of price, and the model determines physical flows and prices simultaneously. Transportation and refining are assumed to be substantially competitive and to exhibit constant returns to scale, that is, they are supposed to operate at a constant long-run marginal cost. The joint supply characteristic of refining operations is captured by the inclusion of an engineering description of refinery technology, with the economic condition that the refiner's margin just cover refinery capital and operating costs. Transportation of crude oil and products among regions is also explicitly modeled, with freight charges determined by costs in the tanker market.

While essentially competitive behavior is assumed in the refining and transportation sectors of the model[2] and by concession holders in oil producing countries,[3] this does not mean, of course, that the countries which export oil are considered pricetakers. Their monopolistic strategy—implemented through the producer cartel, the Organization of Petroleum Exporting Countries, OPEC—is represented by the level of the duty of oil exports.[4] Different levels of this duty (which was increased from approximately $1.85 per barrel in the Persian Gulf in September 1973 to $7.00 by December) imply, through the model solution, different levels of OPEC exports and revenues. One purpose of the model is to investigate the consequences of monopoly behavior by oil exporters through simulation of the effects of changes in the export tax.

The six regions of the model are the United States, Canada, Latin America, Europe, the Middle East and Africa, and Asia. The Communist world is excluded. This set of regions was chosen as the smallest one that adequately represents the important market areas of the world, so that the solution time and cost are minimized and the effects of varying many parameters can be economically simulated.

The model uses five commodities. These include four refined products—gasoline, kerosene, distillate fuel, and residual fuel. The rationale for this breakdown is that different kinds of consumers purchase each fuel type, so the demand functions should be disaggregated. The middle distillates, which are sometimes combined, were divided here to distinguish between the jet fuel and the home heating markets. Crude oil is the fifth commodity.

---

[2] Zannetos [57] argues that the history of freight rates supports the existence of competitive behavior in the transportation sector, and de Chazeau and Kahn [12] similarly contend that the U. S. refining sector is competitive. Adelman [3] shows similar results for the European refining industry.

[3] Adelman [3] presents evidence that the concentration of producers of oil is sufficiently low that this is probable, and that crude price weakness in the 1960s represented the effect of entry by independent producers pushing the marginal return on oil production to competitive levels. From this and the previous note, we conclude that the assumption of competitive behavior, where necessary for modeling the world oil market, is an appropriate working hypothesis.

Of course, this is not a foregone conclusion. One can find arguments (U. S. Congress [54], Tanzer [50]) that the industry is in fact effectively collusive. Although the implications of this change in assumption have not been explored here, the criterion for solution of the model could be changed so as to maximize profit in any one sector. Problems of modeling potential entry and other oligopolistic considerations would then arise, however, so the generalization to noncompetitive behavior is not straightforward.

[4] The financial arrangements for removing oil from exporting nations are actually more complicated than a simple per barrel charge; for an account see Mikdashi [37]. However, since posted prices, which determine income taxes, are set independently of market conditions, and other financial benefits to host countries, such as access to a portion of production of output as participation oil, are related to the volume of production, it was considered adequate to represent the financial obligations of producing companies as a specific export duty.

Trade publications, such as the *Petroleum Intelligence Weekly* in early 1974, described OPEC decisionmaking processes as focusing on the per barrel realization to the host country from production, and then choosing a posted price to correspond to the desired realization. This phenomenon supports the use of the per barrel charge as the appropriate policy variable of OPEC governments.

In the operation of the model, the exogenous variables are the supply and demand equations in each region, the technology of refining, including the cost of capital goods, and the cost characteristics of transportation among regions. Government policy variables are also exogenous, and an important use of the model is to investigate the effects of these policies. They include trade restrictions such as tariffs or export duties, and domestic measures such as excise taxes, environmental restrictions, refinery subsidies, and attempts to shift the supply curve through tax measures or direct drilling regulations. Given the array of exogenous variables, the model has enough information to determine the levels of consumption, production and price for each commodity in each region, the pattern of world trade flows, and the refinery capital structure and output in each region.

☐ **Relation to existing literature.** Several antecedents to this kind of model can be found in the literature. A general survey of techniques for formulating and solving multimarket general equilibrium models is given in Takayama and Judge,[5] most of whose examples come from work done on the agricultural sector. A pure econometric model of the market for an internationally traded commodity, copper, is given by Fisher *et al.*[6] The purpose of that work—forecasting and policy simulation—is similar to ours. A detailed engineering econometric model of the U.S. oil industry has been constructed by Adams and Griffin.[7] This model is similar to ours in its formulation, but takes crude prices and product demand as exogenous. It therefore lacks the desirable property of the simultaneous determination of the market equilibrium. There are no previous examples of formal world oil models of this kind, although Adelman[8] verbally sketches a process of global oil price determination which is consistent with the assumptions embodied in this model.

There is a large industrial organization type literature on the structure and organization of the international oil industry. Outstanding examples of such work are Adelman, Hartshorn, and Stocking.[9] These books concern themselves with the organization and behavior of the integrated major oil companies, the independents, and the oil producing nations. They include, as well, valuable institutional information on the formal concession agreements between oil producing companies and host countries. There is also a descriptive literature on the energy policy of oil consuming nations, and its relation to the price and consumption of oil products. McDonald describes these policies for the United States and Gordon for Europe.[10]

Our model is highly complementary to these studies, since it provides a formal framework for analyzing the quantitative effects of changes in the underlying economic conditions of the market, while the existing literature gives only qualitative conclusions. We sacrifice, of course, much of the richness of the actual fabric of the market, since all institutional details must be abstracted into mathematical terms. In return, we gain quantitative precision, and the ability to handle many variables simultaneously in a manner that is guaranteed to be internally consistent. Both approaches are valuable in analyzing the world oil market.

---

[5] In [49].

[6] In [16].

[7] In [1].

[8] See [3], chapter 5.

[9] In [3], [23], and [47], respectively.

[10] In [36] and [18], respectively.

There is a second body of literature on world oil which is more similar to our work. This is a set of projections of the world energy situation for various future years, which project consumption, supply, and trade flows for various regions. One such study was done by the OECD,[11] and its findings have been updated by Schurr *et al.* and by Darmstradter.[12] These investigations project demand for energy as a whole, both globally and by demand sector (transportation, industrial, residential and commercial, and electric utility), and then confront those figures with assessments of available supplies. Projections along similar lines have been made by oil companies and by industry-government groups.[13] All these studies have merit to the extent that they carefully consider the relation of energy use to the sectoral composition of GNP, and make explicit the assumptions about technical change used to modify historical energy-output relationships.

Unfortunately these studies have two problems in their suitability for the uses for which the world oil market model is designed. The first is a lack of simultaneous determination of prices and quantities. The studies invariably employ historical energy-GNP and composition of energy use relationships for their demand projections. Their underlying assumption, then, is that future price trends will reflect historical ones. But such developments as the new-found strength of the producer cartel, and the increasing stringency of environmental regulations, make this assumption inappropriate. Under these conditions, it is vital that a forecasting method take into account the reaction of consumers of energy to price changes, and the implied equilibrium consumption and price which result from these fundamental economic relationships. The model described here does exactly that.

A second difficulty with existing projections is that there is no simple way for the reader to know the effects of a change in one of the assumptions. In a market as complicated as international oil, a forecasting tool can only be valuable if it can be adapted to the beliefs of the user. Thus, the embodiment of our world oil market model in a terminal operated computer program is very important, for it has on-line capacity to generate forecasts consistent with any set of assumptions. A change in one assumption has repercussions through all the endogenous variables in the model, and the nature of the model guarantees that all the economic relations in the model are satisfied when the implications of a change in underlying factors are investigated. This flexibility is not available in currently published forecasts, for there is no way to modify the conclusions reached as a result of any change in assumptions.

---

[11] In [40].

[12] In [44] and [11], respectively.

[13] See, for example, Shell Oil Company [45] and the National Petroleum Council [38].

One disadvantage of our model in relation to these studies is the relative lack of scope and detail. We have modeled only the oil market, and not other important energy sectors such as coal and natural gas. There is also no breakdown in the demand for oil by consuming sectors; instead, we assume that the response of all users to price and income changes can be captured in one demand curve. These considerations mean, of course, that further research on the world energy market along quantitative lines is needed.

☐ **Mathematical formulation of the model.** A more explicit description of the various components of the model, and how they are combined to generate equilibrium prices and quantities, will now be given. It is first necessary to define some integer parameters:

$NR =$ the number of world regions represented in the model
$N\emptyset =$ the number of physically distinguishable commodities (includes crude oil types and refined products)
$NP =$ the number of economic commodities $(= NR * N\emptyset)$.

The basic output of the model is a set of two vectors, η and d, each of length NP. They contain the equilibrium price and net supply, respectively, of each commodity in the model. That is, η contains the price of each commodity (either crude or refined product) at each region, and d contains the net supply of each. For convenience, these vectors are arranged as NR segments of length $N\emptyset$; that is, the first $N\emptyset$ entries in each vector contain price and net supply of each commodity in the first region, the next $N\emptyset$ entries have that data for the second region, and so on. By convention, all prices are nonnegative, net supply of crude oil is nonnegative, and net supply of (demand for) refined products is nonpositive. We shall now detail the economic relationships which are sufficient to determine the equilibrium values of d and η.

The supply of crude and the demand for products in each region are postulated to depend linearly on all prices in the world market:

$$d = a + A\eta, \tag{1}$$

where a is an NP-length vector, and A is an $NP \times NP$ matrix, of supply and demand parameters. As the intercepts of the net supply functions, the elements of a are partially determined by world income levels and income elasticities of demand for oil products and the existing reserves of crude oil in the ground. The diagonal elements of A are easily interpreted as own-price slopes, and the off-diagonal elements as cross-price slopes. In the case of refined products, we expect the cross-price effects between different regions to be zero, and the magnitude of these effects between different products in the same region to depend on the breadth of definition categories. For example, the cross-price effects between gasoline and residual fuel oil are likely to be zero, but those between high and low sulfur residual fuel should be substantially positive. We have by now accumulated considerable evidence on the nature of the demand functions for refined oil products which we shall present below. Assumptions about supply functions will also be presented below. For convenience in solving the model, equation (1) is rewritten as

$$\eta = Rd + r, \tag{1a}$$

where $R = A^{-1}$ and $r = -A^{-1}a$.

The technology of refining and transportation is represented by a matrix T, with NP rows and NT columns. Each of the NT columns of T contains the input-output coefficients of a single activity. By convention, an input into an activity is represented by a positive number, an output by a negative one. Associated with T is an NT-length vector k, each element of which contains the cost of operating the associated activity at the unit level.

Refining activities input one barrel per day of crude to produce a given slate of products, and refining can (but need not) occur in any region. The typical refining activity thus has nonzero elements only in the entries corresponding to crude and products for one region. There is a $+1$ in the entry corresponding to a crude type, and there are negative fractions in the entries corresponding to refined products. Each region, of course, can produce several kinds of slates, including gasoline oriented product mixes corresponding to American demand patterns, or more heavily fuel oil oriented European slates. The element of k corresponding to a refining activity holds the full capital and operating charge associated with operation of the slate at a level of one barrel per day.

Other columns of T represent trade activities. In the model, shipment of any crude or product is possible between any two points, though most of these shipments will not be profitable. A typical trade activity contains a $+1$ in the row representing the traded commodity in the exporting region, a -1 in the row representing that commodity in the importing region, and zeros elsewhere. The freight charge for moving one barrel per day between the two regions is part of the element of k associated with any trade activity. In the long-run this freight charge is the full capital and operating cost for a barrel per day of capacity in a tanker of optimum size for the given route. Also included in k are any export duties or import tariffs levied on the trade flow, for these are real costs to potential shippers.

The economics of a long-run model dictates that no competitive activity can make a positive profit, for capital would shift into the activity if that were the case. This set of NT constraints can be written as

$$-(T'\eta) \leqq k, \tag{2}$$

because $-(T'\eta)$ is a vector of the gross revenue generated by operating any activity at the unit level, and k is a vector of the capital and operating charge incurred by the operation.

Finally, there is an NT-length vector t, whose elements contain the level of operation of each of the activities. This vector is constrained to be nonnegative. The material balance equation of the model is

$$d = Tt. \tag{3}$$

This assumes that if any commodity is freely disposable, an appropriate slack vector exists in T. Since in competitive equilibrium only activities which are breaking even will be in operation, we derive the complementary slackness equation

$$(k + T'\eta)'t = 0. \tag{4}$$

Typically, of course, most elements of t will be zero. The unknowns in this set of relations are η, d, and t. It can be shown that the relations above are indeed sufficient to determine unique values of η, d, and t, and also that they can be interpreted as the first-order conditions of a quadratic programming problem. The reader is referred to the Appendix for a discussion of this.

Bringing together the results presented above, we have the following equations and inequalities to determine the equilibrium position of the world oil market:

$$\eta = Rd + r \tag{1a}$$

$$-(T'\eta) \leqq k \tag{2}$$

$$d = Tt \tag{3}$$

$$(k + T'\eta)'t = 0 \tag{4}$$

plus the nonnegativity constraints:

$t \geqslant 0$

$d_i \leqslant 0$ if i corresponds to a refined product

$d_i \geqslant 0$ if i corresponds to crude oil.

2. The refining sector

■ An important part of the economic model of the world oil industry described in the first section is the representation of the crude oil refining sector. This sector, corresponding to parts of the matrix T and the vector k, allows crude oil to be transformed into refined products at a given cost. This section describes in more detail the nature of the refining sector of the model, and gives the derivation of the parameters used in it.

As was pointed out in the first section, we assume that the refining industry is competitive. In the long run, this means that the gross margin of refineries, defined as the wholesale value of production less crude oil cost, must exactly equal operating costs plus normal profits on refining equipment. These normal profits include depreciation, income taxes, royalties and fees, and the competitive tax-free rate of return, and are in large part determined by the construction cost of the equipment. Were the refiner's margin less or greater than this, capital would either move out of or into refining, and a long-run equilibrium would not exist. So, in the long run, the cost of construction of refinery equipment is one of the determinants of petroleum product prices.

☐ **The economics of refining.** Petroleum refining is an example of a variable proportions joint product industry.[14] The main input, crude oil, can be converted into different slates of refined products, and the slate chosen by any refiner depends on relative prices of the outputs. The cost of each slate is different in the long run, since some output patterns (e.g., gasoline-intensive ones) require more capital equipment per barrel of output than others. These costs, of course, are just the elements of the vector k in the oil model.

---

[14] A statement of the theory of the joint product firm is given in J. M.

In the long run, the costs of different patterns of refinery output partly determine the world-wide distribution of refining activity. There is a "natural cut" of a barrel of crude oil; namely, that which minimizes refinery capital costs This is the output of a topping plant, i.e., one which merely distills a barrel of oil into its natural components. Further, or downstream, processing of these components to meet demand patterns which vary from the natural cut results in higher costs incurred per barrel.[15] However, since products as well as crude oil can be transported, oil can be refined in one location and consumed in others. The long-run cost minimizing nature of the world model will determine the equilibrium spatial pattern of refining.

However, since this spatial pattern of refining is important to national governments for security, balance of payments, and environmental reasons, they may take steps to change it. Policy options for governments include subsidies or taxes on refineries in their countries, or outright regulations on the amount of capacity that may be built. The effect of these policies can be included in the model, so the final long-run structure of world refining is a combination of purely economic (demand structure, transportation and construction cost) and political factors.

☐ **The history of linear programming refining models.** There is a long tradition in economics of using linear programming (LP) models to represent refinery construction and operation options. Pioneering work was done by Manne and Symonds.[16] Manne's early work concentrated on maximum physical output attainable from a fixed capital stock, by using a model which represented the U.S. refining sector as being located at one point. Marschak[17] extended this model to one depicting the United States as a four-region area, and his model included various transportation activities. These works were particularly concerned with finding the maximum possible output of jet fuel in the short run for national security purposes, and thus did not try to simulate industry behavior. Nonetheless, their approach to the representation of production possibilities by the refinery sector is still the one used today.

More recently, Bonner and Moore Associates[18] used several LP refinery models to predict the cost of lead reduction and hydrocarbon emission reduction in motor gasoline, assuming a fixed

---

Henderson and R. E Quant's *Microeconomic Theory* (New York: McGraw-Hill, 1958)

[15] For a description of the physical processes that occur in refining, the reader can consult, in increasing order of technical difficulty, Manne [34], Hengstebeck [24], and Nelson [39].

[16] In [33] and [48], respectively.

[17] In [35]

[18] In [7].

pattern of demand. This study disaggregated U.S. refining activity into four regions, and recognized the existence of refineries which were below current best practice in their existing capital stock. The work was medium run in nature, because it also took into account optimal refinery expansion patterns as demand grew exogenously. Adams and Griffin[19] used a modification of the same model, regionally reaggregated, to simulate the behavior of the U. S. petroleum refining industry under competitive assumptions. This study showed the ability of the LP model to predict accurately prices in the past, but it assumed that demand levels and crude prices were exogenously determined. In an earlier work, Griffin[20] used the same model to generate short-run average cost curves for the U. S. petroleum refining, which were used to derive a theoretically satisfactory measure of capacity. In summary, LP descriptions of technology have proven to be acceptable for modeling the behavior of the petroleum refining industry

The LP model used here is similar to the latest generation of models referred to above.

☐ **Technical specifications of the refining model.** As outlined in the first section, the refining sector of each region is represented by a set of columns in the matrix T. Each column contains the input-output coefficients of one possible refinery slate in the region. By convention, inputs are positive, outputs are negative, and the activity description is scaled so that the unit level of operation of the activity uses one million barrels per day. Associated with each of these columns of T is an element of the vector k containing the cost of operating the slate at the unit level.

In the five-commodity version of the model (gasoline, kerosene, distillate fuel, residual fuel, and crude) typical refinery slates contained in T might include these:[21]

$$\begin{bmatrix} -0.50 \\ -0.10 \\ -0.20 \\ -0.05 \\ 1.00 \end{bmatrix} \qquad \begin{bmatrix} -0.15 \\ -0.05 \\ -0.30 \\ -0.35 \\ 1.00 \end{bmatrix}$$

The first slate uses one million barrels a day crude input to produce 50 percent gasoline, 10 percent kerosene, 20 percent distillate, and 5 percent residual, with the other 15 percent of the barrel assumed to be refinery waste or output of products not specified in the model, such as lubes, asphalt, or liquefied petroleum gases. It is a typical U. S. refining slate, and its long-run cost of operations (which is the corresponding element of k in the long-run version) is about \$1.00 per barrel. The second slate shown above, typical of European output patterns, is much closer to the natural cut of the barrel and would typically cost about 35¢ per barrel.

---

[19] In [1].

[20] In [20].

[21] These vectors are actually subvectors of the columns T, containing only those elements referring to one region. The other locations in the given columns contain zeros.

In a simulation of the model, a number of such columns would be available for operation in each region, and the totality of refining options is the set of convex combinations of the columns. That is, the output pattern of a region can be any convex combination of the prespecified vectors in T, and the refiner's margin in that region must be just high enough to cover the same convex combination of costs. For example, a region might produce a refinery output pattern corresponding to one-half of both of the vectors shown above. Then, its output slate would be

$$\begin{bmatrix} -.325 \\ -.075 \\ -.250 \\ -.200 \\ 1.00 \end{bmatrix},$$

and the refiner's margin necessary to capture full costs would be 67½¢ per barrel.

This representation of each region's output options is a simplification of the actual technology of refining. In generating the cost associated with each slate, we use a linear programming model of refining technology, which chooses a structure of refinery capital stock that minimizes the long-run cost of producing each slate. There are two problems involved in using the set of output slates and associated costs as a complete representation of refining technology. The first is that in minimizing cost in the LP model, both the cost of crude and the cost of capital equipment are parameters, so that the optimal structure of capital for producing a given slate depends on which crude price is used. The price used, of course, may not be that which is endogenously determined in the final solution of the model. A second consideration is that the cost of a convex combination of output vectors will typically be less than the same convex combination of individual vector costs. This is because the capital equipment patterns most appropriate for the combined slate are not necessarily used in producing any of the vectors which combine to form the new one.

The overwhelming virtue of using a small number of preset output slates rather than a full refinery representation in the world oil model is the reduction in problem size, and thus solution time. The LP model used to generate the preset slates (also called "extreme points") contains 57 rows and 173 columns, and in a 16-region model this would mean that the refining portion of T would contain 912 rows and 2768 columns. Assuming only five products are represented in the world model (as shown in the vectors above) and that twenty extreme points are used to represent each region, this part of T is reduced to 80 × 320, which saves storage space by a factor of one hundred. Aside from this paramount reason for reducing the size of the model, there are considerations which indicate that the simplifications used will not materially affect the final solutions. Concerning the first problem, substitution of equipment for crude is only feasible through a small range, and new extreme points can anyway be easily generated using differing values for crude prices. A dense enough choice of extreme points will alleviate the second problem, and experimentation has shown that a small number, such as twenty for a five-product model, is indeed adequate to accurately represent costs.

The remainder of this section will describe the LP model underlying the generation of refinery slates and costs used in the *T* matrix. The model is based on an LP representation of refining technology supplied by Bonner and Moore Associates.[22] Each column of the LP model represents one possible refinery activity, including the operation of capital equipment and the blending of intermediate streams into final products. The rows represent the constraints on use of equipment and intermediate streams, the output requirements of final products, and constraints on final product specifications. The cost coefficients represent the long-run capital and operating costs of each kind of equipment and the prices of purchased raw materials (including crude). This LP representation was used to construct a smaller (57 x 173) LP model.

The LP model contains a more detailed refined product and crude oil breakdown than that used in the simulations reported here. In refined products, gasoline can be distinguished by octane level (premium vs. regular) and lead content, and residual fuel oil can be distinguished by sulfur content. For simulations of the version of the model reported here, consumption of these products was assumed to occur in fixed proportions, which were relatively stringent with respect to environmental impact. Namely, gasoline output was assumed to be unleaded, 70 percent regular and 30 percent premium, and resid was assumed to be one-half low sulfur (0.5 percent) and one-half high sulfur (2 percent). In the larger LP model, crude oil is also broken down into subcategories by API gravity and sulfur content. For these simulations, all crude was assumed to be a world average, about 32 percent API and 1 percent sulfur content.[23]

An important aspect of the LP model is the breakdown of refinery capital stock by process type. Table 1 lists the kinds of capital equipment explicitly included in the model, as well as the per barrel charge associated with maintaining one barrel per day of capacity of the equipment. This charge includes the operating cost, plus the normal profit associated with ownership of the capital goods. These figures were supplied by Bonner and Moore, and are derived by them by using their assumptions about operating costs, royalties and fees, economic life of equipment, and capital recovery factors appropriate for the refining industry. In practice, however, the necessary annual return to a piece of capital equipment varied little from 30 percent of the purchase price, which would represent a 10 percent operating cost, and a 10 percent pure rate of return with a 50 percent corporate profit tax.

---

[22] Financial support for acquisition of this data was provided by the Energy Policy Project of the Ford Foundation, which supported the first stages of the development of the model.

[23] Data on the gravity and sulfur content of the world's crude oil reserves were found in the *International Petroleum Annual* for 1972. For purposes of finding a world average, only the largest reservoirs were used, so the specifications are approximate.

TABLE 1

REFINERY PROCESS UNITS INCLUDED IN LP MODEL

UNIT	$/DAILY BBL INVESTMENT COST (1972 $)	¢/DAILY BBL TOTAL LONG-RUN COST (1972 $)
CRUDE DISTILLATION	$175	13
VACUUM DISTILLATION	151	11
FLUID CATALYTIC CRACKING	661	54
HYDROCRACKING	1262	107
RESIDUAL HYDROFINING	237	18
MOTOR GAS REFORMING	804	72
GAS OIL HYDROTREATING	158	12
COKER	649	49

An examination of the table shows that the kinds of equipment necessary for intensive gasoline production, such as cracking and reforming units, incur significantly higher per barrel charges than simple topping equipment. This explains the higher refiner's margin needed in U. S. refining. In simulations done with the world oil model, we have assumed refinery charges to be the same in different regions of the world. This assumes, of course, that the net effects of equipment vintage, labor and other operating costs, construction costs, costs of capital, and tax laws cancel out in every area. An important improvement of the model would be to take these factors explicitly into account.

We now present some salient results from the LP model, and thus indicate the essential quantitative nature of refinery economics as used in generating solutions of the world oil market model. The most striking feature of refinery economics is the increased cost per barrel of producing more gasoline oriented product slates. Table 2 shows the capital structure and cost of refineries which

TABLE 2

THE COST AND CAPITAL STRUCTURE OF DIFFERENT REFINERY SLATES

SLATE	ATMOSPHERIC DISTILLATION (M BBL /DAY)	VACUUM DISTILLATION (M BBL /DAY)	CRACKING (M BBL /DAY)	REFORMING (M BBL /DAY)	H_2 TREATING (M BBL /DAY)	INVESTMENT/ DAILY BBL (1972 DOLLARS)	¢/BBL (1972 DOLLARS)	GASOLINE	KEROSENE	DISTILLATE	RESIDUAL
1 ASIA	1000	80	48	186	148	456	35	0 17	0 09	0 17	0 42
2 EUROPE	1000	200	154	129	214	548	42	0 17	0 09	0 25	0 34
3 INDIA	1000	352	189	5	406	565	44	0 09	0 17	0 34	0 26
4 SOUTH AMERICA	1000	215	166	229	226	610	51	0 26	0 09	0 17	0 34
5 UNITED STATES	1000	454	561	260	275	1364	114	0 51	0 09	0 17	0 08

are optimally constructed for producing the slates shown. The crude used here is an approximate world average. The slates shown are named for the region whose output they most closely resemble. As we go from a simple distillation of oil which approximately represents Asian consumption patterns to one which is typical of U. S. refinery output, we see an increase in the complexity of the refinery and the associated cost of building and running it.

3. The demand sector

■ An important part of the world oil model is the set of demand and supply equations

$$\eta = Rd + r. \tag{1a}$$

Here we shall present evidence to justify the parameters used in the matrices R and r for the simulations presented in the last section. This discussion will be restricted to elements of R and r which refer to the demand for products; assumptions about supply will be given in the next section. Since the results referred to here have been reported elsewhere,[24] along with a review of the relevant literature and a discussion of the statistical problems involved in estimation, only a brief discussion of the work will be given here.

☐ **Framework of the study.** The evidence on demand for refined oil products presented here combines time series data from several OECD countries to estimate oil consumption as a function of oil price and income. This approach is valuable in determining the behavior of energy users, because the variation in income and price over countries is greater than the variation in aggregates over time. It requires that the structure of demand be the same in different locations, but this powerful assumption allows the econometric techniques to use all available data to determine those responses to income and price which are in best agreement with the evidence. Some effects of structural differences between countries can also be captured, through the inclusion of additional variables (temperature, population density) or dummy variables. Previous international cross-country studies, such as those by Houthakker and by Goldberger and Gamaletsos,[25] support the hypothesis that patterns of demand in different countries can be usefully related to variations in prices and incomes.

The equations were estimated from annual time series for a number of OECD countries. The income terms have all been deflated to 1963 constant value home currency figures, then converted to U. S. dollars at 1963 exchange rates. Prices have been similarly deflated and converted. Left-hand variables are either in per capita terms or deflated by income level as an *ad hoc* adjustment for heteroscedasticity.

The results are based on the following dynamic flow adjustment model, in which demand for an energy product depends on underlying economic factors (such as price and income), but that the effects of these factors is spread out over time. The model has been successfully used before in demand for energy studies, such as

---

[24] In [27].

[25] In [26] and [17], respectively.

Balestra and Nerlove.[26] A description of this model and its statistical properties in the context of the general distributed lag problem is found in Griliches.[27]

We begin with a level of *desired demand* for a given energy product (q^*), which depends on income (y) and price (p):

$$q^* = f(y, p). \tag{5}$$

(Desired demand may depend on other factors as well, such as temperature, but these are momentarily ignored for clarity of presentation.) We specialize this equation to the log-linear form:

$$q^* = \alpha\, y^{\beta}\, p^{\gamma}. \tag{6}$$

The actual level of demand in one year, however, is not necessarily equal to the desired level. In particular, *actual demand* adjusts toward desired demand with a lag, so that this year's consumption is a function of this year's economic variables and last year's consumption. The adjustment process is represented as

$$q/q_{(-1)} = [q^*/q_{(-1)}]^{(1-\lambda)}, \tag{7}$$

where $q_{(-1)}$ is last period's consumption. This lagged adjustment of actual demand to desired demand is the result of (1) an existing stock of equipment which uses a specific form of energy at a specific efficiency, and which cannot be replaced immediately and (2) an unwillingness by consumers to view price and income changes as permanent until they have continued for some time.

Substitution of (8) into (6) generates the estimating equation:

$$\ln q = \ln a + (1-\lambda)\beta \ln y + (1-\lambda)\gamma \ln p + \lambda \ln q_{(-1)}. \tag{8}$$

In this equation, β and γ are long-run elasticities of demand, and $(1-\lambda)\beta$ and $(1-\lambda)\gamma$ are short-run elasticities.

☐ **Statistical implementation.** Several interesting statistical problems arise in the estimation of these demand curves. One is the problem of multicollinearity, or correlation among the explanatory variables. In recent years, real prices of energy have been falling, while income has been rising. In the cross section of countries, low oil prices are associated with high per capita income. Since high income and low price affect consumption in the same way (positively), there may be some ambiguity in interpretation of the data—did consumption rise because price fell or because income rose? This ambiguity, if it exists, will be reflected in relatively high standard errors of our estimates of price and income elasticities. However, the use of time series for different countries tends to reduce the problem of multicollinearity.

Another problem we have is that of left out price variables.[28] In economic theory, demand for a product or factor of production is a function of its own price, income, and the price of all other goods and services. Our model takes some account of this in that prices are deflated by either the CPI or WPI; technically, this means we assume all other prices increase at the same rate, the general rate of inflation. One price of particular importance in the fuel oil equations is the price of coal. Consistent time series of coal market quotations were not found, although there is some evidence[29] that the overall price increase of coal was comparable to that of the general level of inflation during the time period. If this is so, our estimates are unbiased. If the price of coal rose faster than the CPI, our price elasticities are biased upward in absolute value, and vice versa.

---

[26] In [5].

[27] In [21].

[28] See Griliches [22] for the effect of left-out variables on the estimates of other parameters.

Another statistical problem arises due to the presence of a lagged dependent variable on the right-hand side of equation (8). Coupled with autocorrelation in the error term, which may result from either the adaptive expectation process,[30] or left out explanatory variables, this will lead to biased estimates of the coefficients. The procedure used here is outlined by Dhrymes,[31] where a proof of its consistency can be found. It assumes that any autocorrelation in the error term can be expressed as a first-order autoregressive scheme. Initial consistent estimates of the coefficients are found by using an instrumental variable (lagged population) for the lagged dependent variable. Then, a consistent estimate of the autocorrelation coefficient is derived, which is used to quasi-first-difference the observation to obtain a second round of consistent estimates. Estimated values of the autocorrelation coefficient were typically around 0.5. As is predicted by econometric theory, ordinary least squares estimates of the equations produced estimates of λ much higher than those obtained by the consistent method.

A final statistical problem is the familiar one of simultaneous equation bias due to the existence of a supply equation.[32] Fortunately, data on the level of excise taxes are available over the sample period, so it could be used as an exogenous variable in a two stage least squares framework. This assumes, of course, that taxes are exogenous, and not affected by wholesale oil prices. The use of taxes to protect coal in some countries in Europe indicates that this assumption could usefully be explored further.

☐ **Results.** The equation for gasoline demand was run on twelve countries,[33] and covered the years 1962 to 1972. Consumption is measured in per capita terms, and the income variable is real per capita consumption expenditures. The full estimated equation also included variables for population density and temperature.

Gasoline

$$\begin{aligned}\ln q = 0.213 &+ \underset{(0.101)}{0.744 \ln y} - \underset{(0.105)}{0.465 \ln p} + \underset{(0.068)}{0.442 \ln q_{-1}}\\ R^2 &= 0.93.\end{aligned}$$

---

[29] See Gordon [18].

[30] See Griliches [21].

[31] In [13].

[32] See Working [55].

[33] Portugal, Italy, Austria, Belgium, Denmark, France, West Germany, Netherlands, Norway, Sweden, the United Kingdom, and the United States.

The implied long-run income elasticity is 1.3, and the long-run price elasticity is —0.82. The figures are significantly higher in absolute value than those found in recent studies of gasoline demand for the U. S. alone,[34] although our sample includes the U. S. One explanation is that the log-linear function is only a local approximation, and that the lower income levels and higher price levels typically prevailing in Europe lead to higher elasticities.[35] At any rate, this is strong evidence for the influence of price on gasoline consumption, other things being equal.

The remaining equations were fitted over a smaller sample of nine countries,[36] for the years 1965 to 1970. In the kerosene and distillate fuel equations, dependent variables were expressed in per capita terms, and per capita consumption expenditures and prices, both deflated by the consumer price index, were used as explanatory variables. This specification assumes that the demand for these products can be captured by variables relevant for residential consumers. In further research, the demand for the fuels should be disaggregated by consuming sector. In any event, during the time period under consideration, residential use did predominate in middle distillate consumption.

Kerosene

$$\begin{aligned} \ln q = 2.36 - \underset{(0.082)}{0.207} \ln y - \underset{(0.186)}{0.173} \ln p + \underset{(0.039)}{0.927} \ln q_{-1} \\ R^2 = 0.93. \end{aligned}$$

The long-run income elasticity here is —2.5, indicating that kerosene is an inferior good, in agreement with common sense. The long-run price elasticity is a rather high —2.0, indicating the existence of good substitutes, no doubt #2 heating oil and coal. These results are intuitively quite plausible when the role of kerosene as heating oil is considered, but the increasing use of kerosene as jet fuel makes them suspect for projection purposes. As jet fuel, the income elasticity of kerosene demand would likely be over one, and the lack of good substitutes should make the price elasticity low in absolute value (although clearly not zero). This appears to be a case where changing underlying structure vitiates the projective power of historical data.

In the distillate equation, a separate dummy variable for each country was included:

---

[34] For example, Houthakker, Verleger, and Sheehan's [28] estimate of the long-run elasticity was —0.25.

[35] However, carrying out a Chow test shows that the coefficients for the United States and for all other countries are not significantly different. The computed value of the test statistic is 0.17, while the critical F value is $F(6,108) = 2.19$. Essentially nothing is lost by disaggregation. However, the U.S. observations included here are for the country as a whole, and when the equation is estimated on them alone, the price variable is not significant. Thus, regional disaggregation, such as that included in the study of Houthakker, Verleger, and Sheehan [28], is needed to pin down the U. S. elasticity.

[36] Japan, Italy, Belgium, Denmark, France, Germany, Netherlands, Sweden, and the United Kingdom.

Distillate fuel oil

$$\begin{aligned}&\ln q = -0.231 + \underset{(0.436)}{1.43}\ \ln y - \underset{(0.193)}{0.388}\ln p + \underset{(0.150)}{0.471}\ln q_{-1}\\&R^2 = 0.90.\end{aligned}$$

These results are quite plausible. The long-run income elasticity is 2.8, indicating that distillate fuel is a superior good. Since its substitutes, coal and kerosene, are much less convenient, this is very likely. A lower elasticity might be expected in the forecast period, however, as (1) saturation of fuel oil heat begins to appear, and (2) an even more convenient fuel, natural gas, comes onstream in some European countries. The long-run price elasticity is —0.76.

In the equation for residual fuel oil, GNP was taken as the income variable, and prices were deflated by the wholesale price index. The dependent variable was deflated by GNP, so the income elasticity is one plus the coefficient on ln y. Again, all uses of resid are lumped in one equation.

Residual fuel oil

$$\begin{aligned}&\ln q = 0.884 + \underset{(0.146)}{0.396}\ln y - \underset{(0.426)}{1.05}\ \ln p + \underset{(0.207)}{0.335}\ln q_{-1}\\&R^2 = 0.96.\end{aligned}$$

The long-run price elasticity is a very high —1.58, reflecting the close competition with coal in this market. The income coefficient of 1.6 in the long-run implies that either residual fuel is superior to other fuels due to convenience factors, or that the sectoral composition of GNP was shifting toward residual-intensive uses over the sample period. Both factors seem likely, as the share of manufacturing output was increasing in OECD countries at this time and the electricity-GNP ratio was rising. The adjustment coefficient, $\lambda = 0.34$, is quite low. This reflects the rapid growth in industrial capacity, and thus the turnover of energy-using equipment, during this time. The use of the WPI as a deflator in this equation may bias the price elasticity estimate downward in absolute value, since the price of labor, an important variable in mechanization decisions, rises much faster than the WPI.

All of the equations presented in this section tell the same story. Price is an important factor in demand for energy, in particular for refined oil products. Blind extrapolations of past consumption trends are misleading, because recent history is a reflection of falling real prices, coupled with price elasticities in the neighborhood of —1. Therefore, oil use grew much faster in recent years than any measure of income. The abrupt recent turnabout in oil prices will sharply curtail demand growth, just as previous price weakness encouraged it. The world's economy is no more "locked into" oil use today than it was locked into coal use in 1950. Relative price changes, as reflected in the historical evidence presented here, will determine the use levels of various forms of energy. The implication of these findings for future developments in the world oil market will be presented in the last section.

Although these results are clear enough for use in a forecasting model, there is certainly room for expanding them. One important step would be explicit modeling of interfuel competition through the inclusion of alternate energy prices. This would enable us to determine how much of the overall price elasticity of oil demand reflects substitution of other forms of energy, and how much pure energy conservation. Another interesting modification would be a breakdown of oil use into demand sectors (residential, industrial, transportation) so that different patterns of demand in these sectors are distinguished. This, we noted, was particularly important for kerosene. Finally, inclusion of more countries, particularly the less developed ones, would test the robustness of the results. As it is now, we must use the elasticities estimated for developed countries in projections with the world model. However, a serious extension of the demand study in this direction would necessitate a major data development effort, and would be a major project in itself.

☐ **The use of demand equations in the model.** For the simulation results reported in the last section, values of price and income elasticities had to be chosen for future years. Based on the regression results reported here and some *a priori* considerations, the following elasticities were chosen. (The same values are used for each region.)

Product	*Price Elasticity*	*Income Elasticity*
Gasoline	−1.0	1.0
Kerosene	−0.5	1.0
Distillate	−0.5	1.0
Residual	−1.0	1.0

The price elasticities, of course, refer to the retail level, so lower values apply to wholesale price changes. World income is assumed to grow at 4 percent in all regions. Some substantial modifications to the regression results have been made for projection purposes. Distillate and residual fuel oil income elasticities have been reduced to unity, to account for the once-and-for-all shift away from less convenient substitutes in distillate consumption, and the once-and-for-all change in the composition of GNP toward energy intensive output. Kerosene elasticities have been adjusted since the use of this product as jet fuel will become more important. Finally, the price elasticity on resid has been reduced in absolute value to account for the possibility of an increasing supply price of its substitute, coal. To test the sensitivity of results to these assumptions, a set of runs is included in which all income elasticities are 1.5. This means effectively that, at unchanged real prices, oil consumption would grow at 6 rather than 4 percent.

The results on the demand for refined products discussed here are given in terms of log-log equations, while the demand sector of the world oil model contains linear demand equations. The log-log (constant elasticity) formulation of the demand curve was found to be best in fitting the actual equations, while the linear demand sector in the world oil model is a convenient approximation to the true demand equations for solution purposes. The rest of this section describes how the linear approximation is derived.

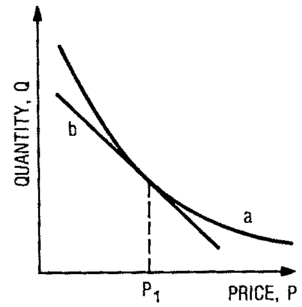
For the four refined product categories we assumed a zero cross-price elasticity of demand, so in the equation,

$$d = a + A\eta, \tag{1}$$

consumption is a function of own price only. An entire linear demand curve can be derived from one point and the slope. The curves used in the world oil model were calculated by the following two-step procedure. A first guess at the 1980 equilibrium price for each region was made, and, by using some assumption about income growth to 1980, a point on the demand curve was found. Then, by using the slope of the demand curve at that point, a linear approximation was made to derive a linear demand curve.[37] This procedure is illustrated in Figure 1. Curve *a* is the true demand curve. P_1 is the assumed price in 1980. Line *b* is the approximate demand curve used in the world model.

FIGURE 1

DERIVATION OF LINEAR APPROXIMATION TO DEMAND CURVE



4. Other sectors of the model

■ This section will describe those subsectors of the world oil model not previously discussed. They are the crude supply sector, the transportation sector, and the retail market sector. These are areas in which no original results have been derived for use in the world oil model, but in which parameters have been adapted from the existing work of others.

☐ **The crude supply sector.** The formulation of the world oil model in Section 1 contains the generalized net supply equation

$$\eta = r + Rd,$$

some rows of which contain the supply price of crude oil as a function of output. This section will explore some of the issues involved in expressing crude supply as a simple function of price, then review existing results on the nature of this function and detail the assumptions made in simulations of the world oil market model.

In the short run, when the stock of capital installed in oil fields (wells, pipelines, etc.) is fixed, there is a textbook-like increasing relation between output and marginal cost.[38] This is typically very flat up to a normal output level, then sharply increasing. There is little argument that crude production is relatively inelastic in the short run, with even large price increases unlikely to make substantial increases in production profitable.

Once one enters the long-run situation, however, things are very different. In the long run, two kinds of activities are carried out by economic agents: provision of capital goods for exploitation of existing oil fields, and exploration for new fields. The optimal development of an existing field is a typical problem in investment theory, since outlays for necessary equipment and services must occur before any output will appear. In addition, the output continues over a long time span, typically fifteen to twenty years.[39] Existing economic studies of the theory of the optimal development of a petroleum reservoir, such as those by Bradley and by Baughman,[40] show that the desired level of capital spending, and thus the total output, is a positive function of future expected prices, as is normal in an investment problem.[41] Thus, with a known stock of petroleum reservoirs, we expect some response of output to price in a time period long enough to vary the capital intensity of production.

---

[37] This can be interpreted as a first-order Taylor series approximation to the true curve.

[38] An example of this relation for the United States in 1965 is given by Steele [46].

[39] A comprehensive list of the most important oil fields in the world, including their current and cumulative output and date of discovery, is given in *International Petroleum Encyclopedia* (1973).

A second mechanism for varying the production of crude oil as price changes in the long run is exploration activity. As was pointed out in work by Grayson and by Kaufman,[42] the determination of the optimal exploration effort is a problem in statistical decision theory. Again, economic theory predicts that exploration activity will increase as expected future prices increase (or as the distribution of future prices moves to the right), so we have a second mechanism by which output will increase with price in the long run.

Unfortunately, the exact specification of both development and exploration activity requires a future distribution of crude oil prices, and the world oil model, static in nature, can only represent supply as function of one price. Therefore, we assume that the oil production sector expects real prices to remain unchanged from their equilibrium value (this would be obviously untrue of money prices) in the long run, and that the supply function represents output for the given year under that expectation. To move significantly away from this simplification would require a detailed model of producer expectation formation, which is beyond the scope of this study.

A third problem in the theoretical consideration of crude supply is the existence of noncompetitive elements in the market. In particular, a large portion of the world's crude oil output is produced in countries which are members of OPEC. These countries do not act as price takers, but they also do not follow a simple joint profile maximizing cartel strategy, which would be to concentrate all production in the lowest cost region, and then arrange side payments among the members.[43] The most desirable representation of OPEC behavior would be in a game theoretic framework, in which members could collude or price chisel depending on their utility functions and their perceptions of the responses of others. In the version of the oil model currently operating, a drastic simplification is imposed. Non-Persian Gulf and North African producers are assumed to maintain their production at current levels regardless of price, and Persian Gulf and North African producers then satisfy world excess demand at export duty plus cost. This amounts to an optimistic assumption from the viewpoint of Gulf and North African producers, for it means their potential competitors do not try to increase their revenues by shading prices when the world market price is high. An extension of the model along the same theoretic lines indicated above would greatly improve the representation of OPEC.

The most relevant empirical evidence concerning the relation of long-run oil supply to price will now be reviewed. A straightforward econometric analysis of the relation of proved reserves to real oil price was done by Erickson and Spann.[44] Using U. S. historical data, they found a reserve-price elasticity of 0.9. Assuming that the recovery ratio is positively related to price, we can use this figure as the minimum long-run price elasticity of production. A different approach to finding the supply curve was taken by Mancke.[45] In a competitive market for leases to produce oil from privately owned land, the rent paid to the landowner would be equal to price less full cost of production. Using data on rent payments to landowners, and knowing prevailing price, one can find the statistical distribution of the cost of production of U. S. onshore oil which determines the competitive supply function. From these data, Mancke estimates the U. S. elasticity of supply to be between one and two. Finally, submissions of U. S. oil companies to the U. S. Cabinet Task Force[46] concerning quantities of oil that could be profitably produced at different prices also imply a price elasticity of around unity.

---

[40] In [9] and [40], respectively.

[41] For a review of the theory of investment, which shows the dependence of optimal capital spending on future prices, see Jorgenson [29].

[42] In [19] and [30], respectively.

[43] See Scherer [43].

In the world oil model simulations reported below, the following assumptions were made about supply. In the Persian Gulf, currently the world's largest producer, unlimited supplies were assumed to be physically available at 25¢ per barrel in 1972 dollars. Evidence for this assertion can be found in Adelman.[47] On top of this real cost to the producer, of course, is the government royalty charge for the extraction of oil from its territory. The magnitude of the royalty is an important policy variable for simulations of the world model. The entire monopoly position of the Persian Gulf oil-bearing nations is summed up in this royalty, since we assume as much oil as is desired can be obtained from the area at royalty plus cost.[48] Other oil producing regions associated with OPEC are assumed to maintain production at current levels regardless of price. As described above, this behavior is a relatively favorable outcome of internal OPEC bargaining from the viewpoint of the Persian Gulf. OPEC behavior is then summarized as: (1) a Persian Gulf royalty is set, (2) other OPEC nations keep production at current levels, and (3) the Persian Gulf is the swing producer. Problems of prorationing of production among Gulf nations are not considered explicitly.

We must also consider the condition of oil supply in the non-OPEC world, meaning effectively the United States, Canada, and Europe. For Canada and the U. S. lower 48, we assumed a decline rate, from 1972 levels, of 3 percent per year at unchanged real prices. For different real prices, a supply elasticity of 1/3 was used. As seen above, this is very low relative to existing estimates, but the possibility of supply bottlenecks of important exploration and drilling equipment until 1980 makes a smaller response than that historically observed seem more likely. New oil supplies, from

---

[44] In [15].

[45] In [32].

[46] See [53], Appendix D.

[47] In [3].

[48] Determination of optimal OPEC policy can then be made by varying the royalty, and evaluating the desirability of each outcome to OPEC according to some utility function.

Alaska and the North Sea, are assumed to be available in quantities of 2½ and 4 million barrels a day respectively, regardless of price. We further assume that Canada and the European North Sea producers will not act as members of OPEC, but will make their oil available at world price in quantities large enough to make their marginal cost equal price. To test the sensitivity of the results to the supply assumptions, we will report an alternate set of simulations in which U. S. and Canadian supply elasticities are increased to 0.67.

☐ **The transportation sector.** An integral part of the world oil model is the shipment of crude and products between regions. These activities are represented in the second part of the T matrix, each column of which contains the input-output coefficients for shipping a given commodity (either crude or product) from one region to another. The corresponding part of the k vector, of course, contains the charge associated with unit operation of the activity, in this case one million barrels a day. For this study, no new work was done on determining the equilibrium level of long-run shipping charges, but forecasts of these for future years were taken from other sources.

One important set of transportation activities is pipelining and coastal shipping in North America. The costs of using these routes in recent years were collected from the U. S. Bureau of Mines and Bonner and Moore Associates[49] and for projection purposes the real costs were assumed to remain the same in the future. Table 3 shows the per barrel charge associated with some of the more important activities.

TABLE 3

COSTS OF INTERNAL NORTH AMERICAN SHIPMENTS (1972 DOLLARS)

ROUTE	COST (¢/BBL.)
U S GULF COAST–EAST COAST	45
U S GULF COAST–MID WEST	25
WESTERN CANADA–U S MID WEST	60

For the long-run costs of ocean transportation of oil, estimates for future years were generated from a model of tanker rates developed by Dr. James Moose of Loomis-Sayles Inc. For a given year, this model computes the per barrel charge associated with transportation along any given route. The principle used in development of this charge is similar to that used in the refining sector, in that the charge must cover operating costs along the route, plus the full capital charge necessary to maintain tanker capacity in service along the route. This capital charge includes depreciation, insurance, and a 10-percent after tax rate of return to equity, the same as in refining. In developing these freight rates, it was not assumed that the cost of tanker construction would rise with the general price level (as was assumed in refining), but that it would be 20 percent higher, relative to other prices, in 1980. Operating costs were assumed to be constant in 1972 dollars through 1980.

Due to large economies of scale in tanker construction, the long-run equilibrium freight rate between two points is very sensitive to the maximum size tanker that can make the voyage. Table 4 shows some of the more important routes in the world model, the assumed maximum size tanker capable of making the trip, and the associated freight rate. Since inflation is assumed in construction costs, these charges are higher than they would be in 1974. An important application of the world oil model would be to simulate

---

[49] In [52] and [7], respectively.

TABLE 4

COSTS OF OCEAN SHIPMENT OF CRUDE OIL
(1980 CHARGES IN 1972 DOLLARS)

ROUTE	COST (¢/BBL)	TANKER SIZE (000 DWT)
FROM PERSIAN GULF TO		
U S EAST COAST	155	60
CANADA EAST COAST	165	60
NORTH WEST EUROPE	100	250
JAPAN	55	380
FROM CARIBBEAN TO		
U S EAST COAST	35	60
CANADA EAST COAST	50	60
NORTH WEST EUROPE	45	250
FROM NORTH AFRICA TO		
U S EAST COAST	70	60
CANADA EAST COAST	85	60
NORTH WEST EUROPE	35	200
FROM WEST AFRICA TO		
U S EAST COAST	75	60
CANADA EAST COAST	90	60
NORTH WEST EUROPE	45	250
FROM INDONESIA TO		
U S WEST COAST	80	150
JAPAN	35	150

the consequences of lower freight rates on world consumption and trade patterns. Construction in the United States of port facilities capable of handling 380,000 DWT tankers would reduce all freight charges by approximately 60 percent, assuming, of course, the loading port was also capable of handling these large carriers.

☐ **Retail markets.** We shall now describe the last piece needed for the world oil model; namely, the bridge between retail prices, of which demand, as described in Section 3, is a function, and wholesale prices, which determine the output of refineries in the region, as described in Section 2. Three components go into the difference between average wholesale prices in a region and average retail prices. These are retail margins, internal transportation costs, and excise taxes. For projection purposes, we assumed that these would stay constant in real terms throughout the forecast period. In simulating the model, different assumptions about these magnitudes can be inserted and the resulting effects examined. For example, a reduction in high European gasoline excise taxes has been proposed by OPEC countries as a way for the effects of their export taxes to be lessened. The impact of such a policy change could be simulated.

An important implication of this constant real cost gap between retail and wholesale prices is that wholesale price elasticities of demand will be less than retail. The magnitude of this effect for the important consuming regions of the world is shown in Table 5.[50]

---

[50] It was necessary to develop estimates of average 1972 tax levels and retail margins for each region. Various government and trade sources were used to develop these figures, including the U. S. Bureau of Mines [51], E.N.I. [14], and Platt's [41]. Due to selective coverage of these sources, data for some of the countries often had to be extrapolated to cover an entire region.

TABLE 5

IMPLIED WHOLESALE PRICE ELASTICITIES OF DEMAND GIVEN TAX AND RETAIL MARGIN ASSUMPTIONS

PRODUCT	RETAIL PRICE ELASTICITIES	UNITED STATES	EUROPE	JAPAN
GASOLINE	1.00	0.49	0.22	0.27
KEROSENE	0.50	0.35	0.15	0.26
DISTILLATE	0.50	0.34	0.15	0.20
RESIDUAL	1.00	0.93	0.88	0.93

For a large percentage increase in wholesale price (such as that implied by the recent OPEC price policy), the reduction in consumption is smaller than that predicted by the retail demand function, the higher the retail wholesale gap. In particular, products with high taxes and retail handling charges, such as gasoline, show much less demand responsiveness to the price of crude than does residual fuel, with its relatively low retail-wholesale margin.

5. Results of world oil model simulations

■ The previous material has discussed the framework and components of the world oil model in a general way. The focus will now be on specific implementations of the model; that is, the specification of input parameters to the model, and the interpretation of the results.

☐ **The problem and our approach.** In the middle of the development of this model, from October until December 1973, a revolution occurred in the structure of world oil prices. The f.o.b. value of a barrel of crude oil at Ras Tanura increased from \$1.75 to \$7.25, on a tax plus cost basis. This price increase was a manifestation of the recently developed strength of the oil producer cartel, OPEC. The world oil model seemed an ideal tool for use in assessing the probable future developments of the world oil market in light of the new circumstances. In particular, two questions seemed crucial. First, what was the effect on the world oil consumption, production, price, and trade structure of various policies set by the OPEC nations? Second, what were the implications of these effects for the stability of the cartel, and thus for the long-run tenability of any policy? The first question is answered directly by several simulations of the model, in which OPEC policy is exogenous. The answer to the second depends on judgments about OPEC member motivations which are outside the scope of the formal economic model.

There was considerable controversy on the subject. Polar views were expressed by oil experts Adelman and Akins.[51] Their positions will be briefly summarized here. Adelman maintained that the cartel was potentially unstable, and thus could be broken up if the oil importing countries would adopt sensible buying policies. One policy he saw as desirable was forbidding the multinational oil companies, whose legal residence is in importing countries, from directly marketing OPEC crude. Thus the importing governments could deal directly with producer governments, and through offers of large, long-term sales contracts could induce price chiselling and the demise of the cartel. Akins, on the other hand, contended that the rapidly rising consumption of oil products in the world gave the OPEC producers an economic stranglehold over the importing countries, since large new quantities of oil could only be produced inside their borders. The dependence of importers on OPEC oil for a commodity necessary to maintain their economic prosperity meant that any producer mandated price increases could only be absorbed by customers, and that the best policy of importers was to use their political and economic influence to forestall price increases.

---

[51] In [2] and [4], respectively.

Clearly each of these observers was making assumptions about economic parameters, and using some sort of implicit model to transform assumptions into outcomes. However, their assumptions and implicit models are not readily available for consistency evaluation or sensitivity testing by the reader of their policy prescriptions, so little objective evaluation of their arguments is possible. With the world oil model described here, however, one has an internally consistent set of outcomes that follow rigorously from the specified assumptions, and one can vary key parameters to identify results that are highly dependent on personal assumptions. With this tool, we hope to analyze more scientifically the future course of oil market conditions, and to expose clearly the logic of the conclusions to the reader. Furthermore, we shall indicate the sensitivity of the results to changes in some of the parameters about which there is considerable uncertainty and dispute today.

☐ **The framework and assumptions.** For this task, the six-region five-product long-run version of the model described above was used. The long-run formulation was appropriate because we are interested in the conditions of the oil market for the next ten years or so, and in this time period long-run economic forces, which induce changes in the composition and size of capital formation, predominate. It is clear that almost any price of oil is possible in the very short run, when net demand for OPEC output is highly inelastic. However, this situation, as politically important as it may be, is not relevant to the actual economic framework in which OPEC must survive, which is longer than a year and in which actual and potential customers can change their behavior with respect to oil use. In particular, we want to know whether any given OPEC policy can be maintained in the face of the substitution by consumers of other products for oil (demand response) and the production of non-OPEC oil (supply response) which will occur as a result of the OPEC actions.

Since a static model is being used, the solution must be found for a given year. 1980 was chosen as the target year for the simulations reported here, on the grounds that the time until then is sufficient for long-run responses to have occurred, and the use of 1980 in other forecasts, such as those described in the first section, makes our results easily comparable with theirs. The disadvantage of a static model is that only the equilibrium values of important variables can be determined, and not their time paths. An important extension of the world model would be an adaptation of it for year-by-year dynamic solution, in which demand and supply functions depend partially on previous prices and other economic variables.

☐ **Results: base case.** Table 6 lists the values of the Persian Gulf export tax in 1972 dollars for which simulations of the oil model,

TABLE 6

RESULTS OF BASE CASE SIMULATIONS

ROYALTY AT PERSIAN GULF ($/BBL)	PERSIAN GULF AND NORTH AFRICA EXPORT VOLUME (MM BBL /DAY)	PERSIAN GULF AND NORTH AFRICA EXPORT REVENUE (BILLION $/YR)	OTHER OPEC EXPORT REVENUE (BILLION $/YR)
0.80	39.3	14.6	4.4
1.75	31.1	22.0	6.8
3.50	21.2	28.1	11.2
5.25	14.9	29.1	15.6
7.00	11.1	28.9	19.3
8.75	7.4	24.4	23.2

given the base case assumptions described above, were done. For each of these values, the table shows the daily volume of Persian Gulf and North African exports, the yearly revenues derived from them, and the yearly oil export revenue going to other OPEC producers. Since the latter nations are assumed to have fixed production levels, their revenue naturally increases with world price levels, and thus with the Gulf tax level. Persian Gulf and North African revenues, however, remain relatively constant throughout the tax range $3.50 to $7.00, as increases in the tax are just offset by decreases in overseas sales. This means that the demand curve of the rest of the world for Persian Gulf and North African crude (which is simply an excess demand curve) has unitary elasticity through a wide range. (Detailed results of all simulations described in this paper are available from the author on request.)

The level of maximum attainable revenues, 30 billion dollars per year, is well below the current yearly proceeds of these producers (50 billion), and certainly below the 100 billion level which can be derived for 1980 by extrapolating recent export growth rates with current prices. It is also a level which would be easy for oil importers to finance relative to expected 1980 trade and income levels. A first result of the model, therefore, is that Persian Gulf and North American producers, like other monopolists, have a definite limit to the profits they can extract from prospective customers. Which of the set of profit maximizing duties will be chosen depends on other factors, which will be considered below. First we must examine the economic factors behind the steep drop in OPEC oil export volume that occurs as the export duty rises.

Tables 7a and 7b show, for selected areas of the world, the production and consumption of oil in 1980 which would occur at different levels of the Gulf export duty. On the consumption side, the market impact of price changes on demand levels is clearly seen. At very high world prices, corresponding to very high export tax levels, consumption in 1980 is predicted to be less than 1972

TABLE 7

CONSUMPTION AND PRODUCTION OF OIL IN IMPORTANT IMPORTING REGIONS IN 1980 UNDER DIFFERENT LEVELS OF PERSIAN GULF DUTY (MM BBL./DAY)

A—CONSUMPTION			
DUTY/ REGION	UNITED STATES	EUROPE	JAPAN
0.80	24.4	19.9	6.8
1.75	22.0	17.5	5.8
3.50	19.1	15.1	4.7
5.25	17.1	13.8	4.0
7.00	16.0	12.8	3.6
8.75	14.8	12.0	3.4

B—PRODUCTION			
DUTY/ REGION	UNITED STATES	CANADA	EUROPE
0.80	8.6	1.3	4.4
1.75	9.4	1.5	4.4
3.50	10.6	1.7	4.4
5.25	11.6	1.8	4.4
7.00	12.2	1.9	4.4
8.75	12.9	2.0	4.4

levels, despite the 4-percent growth in real income postulated for the period. Detailed results of these simulations show that the most affected product is residual fuel oil. This is because the retailing cost and excise tax levels associated with resid are relatively low, so wholesale price levels, which are directly linked to crude prices, are a major portion of the cost to the final user. However, there is some impact on consumption of all products as a result of price changes, for reasons detailed in Section 3.

There are also noticeable changes in the production levels of non-OPEC regions in the 1980 simulations as seen in Table 7b. First is the availability of Alaskan and North Sea crude oil at a level of 6½ million barrels per day. In addition, the much higher cost of crude in the United States and Canada relative to historical levels induces an increase in production from conventional sources.[52] This increase, of course, must be set against the 3-percent decline which was assumed to occur at unchanged real prices.

The relatively sharp drop in sales volume as the tax increases has definite implications for the sustainability of any price policy by Persian Gulf and North African producers. For instance, at the upper reach of the revenue maximizing range, $7.00, exports are only 11 million barrels per day. This figure compares to the 1972 level of 20 million, of which 9 million came from North Africa and Iran alone. Allocation of this reduced level of exports among the producing countries may prove difficult, especially since in the immediate future, when demand and supply are highly inelastic, the countries will be enjoying revenues well in excess of the long-run possible maximum. The existence of some producers, such as Saudi Arabia, which are currently generating revenues well above what they can usefully spend, is a partial cushion against this effect, since production can be reduced there at presumably no loss in utility to the government. However, the reduction in export volume predicted at the $7.00 export tax in 1980 is 125 percent of Saudi Arabia's 1973 peak output, which would prove difficult to allocate among the region's exporters.

Similar considerations apply to the viability of a $5.25 export tax. The export volume implied by this tax, 15 million barrels per day, is again lower than current levels, although the reduction is less drastic. The $3.50 tax, however, leads to exports of 21 million, and should be easy to maintain. On the basis of this simple oligopoly analysis, a price between five and three-and-a-half dollars, but trending toward the lower end, is the maximum sustainable in the long run.

---

[52] U. S. consumption appears to be higher than production at an $8.75 duty in Tables 7a and 7b. This is because U. S. production of natural gas liquids is assumed to be at a level of 1.9 million barrels per day in the model.

One objection to these conclusions is that Persian Gulf and North African producers would have little difficulty allocating reduced levels of output, since they perceive remaining oil in the ground as having high value. However, the remaining oil is only more valuable than an investment made with the proceeds of its sale if its price is rising faster than the effective return on investment. At very high revenue levels, this effective return might indeed fall toward zero, but at the revenue levels implied by this model, all of the income could be usefully spent by the producers. Therefore, the pressures to increase production will be the same as those facing any oligopoly.

In summary, analysis of economic conditions in the world oil market has enabled us to find the implications of various policies undertaken by an important player in the game, the Persian Gulf and North African OPEC region. On the basis of this, we can predict that current price levels are unlikely to be maintained in constant dollars, and that even an export duty as high as $5.25, which generates the same revenues as a range of lower prices, will be difficult to maintain, given the production allocation problem. Since we do not have a formal model of cartel behavior, this is all that can be said strictly as a result of the economic modeling done.

A final aspect of the runs discussed here is of interest, namely, the cost to the developed world of oil imports at various price levels. Table 8 shows these figures for the United States, Europe, and Japan. First we note that yearly sums do not exceed 50 billion dollars per year at any price, and from this result, we can predict that the oil market will not cause any permanent world balance of payments problems. A second, somewhat more startling result, is self-sufficiency of the United States in oil, after thirty years, if the Persian Gulf duty is set at an extremely high level. But this is in accord with the simple economics of the oil market. In the prewar period, U. S. oil was cheap by world standards. This situation came to an end with the onset of cheaper Middle East oil, so U.S. consumption from its own production was displaced by imports. A high enough export duty, though it is a payment to governments

TABLE 8

BASE CASE COST OF OIL IMPORTS
(BILLION $/YR)

DUTY	COST OF IMPORTS TO			TOTAL PAYMENTS TO OPEC
	UNITED STATES	EUROPE	JAPAN	
0 80	13 0	11 7	4 9	19 0
1 75	13.6	14 4	6 4	28 8
3 50	12 4	18 5	7 9	39 3
5 25	8 7	22 0	9 1	44 7
7 00	5 1	24 9	10 0	48 2
8 75	-0 2	27 5	10 7	48 0

and not to nature, would just as effectively make U.S. oil cheap again. However, the results in Table 6 show that the $9.00 Persian Gulf price, which is needed for self-sufficiency to occur solely as a result of market forces, is not likely to persist until 1980. Although OPEC could make U.S. oil exportable if it chose, it is naturally not in its interest to do so.

☐ **Alternate case: high supply elasticity.** As explained in Section 4, the supply elasticity assumed for the United States and Canada in the base case simulations, 0.33, is very low relative to existing econometric estimates. This is because there are considerations which indicate that the historic supply response, on which the econometric literature is based, may overpredict the 1980 response. These include uncertainty about the time lag involved in expanding output, especially given constraints on availability of materials used in field exploration and development, and existing judgmental forecasts, such as the NPC's, which imply low potential for expansion of U.S. lower 48 production. The 0.33 figure was chosen as a very conservative estimate, so that the base case results would not depend heavily on an uncertain parameter. We were interested, however, in the oil market conditions if future North American supply reacted in a manner more similar to the past.

Therefore, a second set of simulations was run for 1980, using the same assumptions as the base case but with a supply elasticity of 0.67. Results of these simulations are given in Table 9. As expected, the outcomes are more adverse for the Persian Gulf and North African producers. Under these assumptions, their revenues hit a clear peak between a $3.50 and $5.25 export duty, and at very high world price levels they are reduced to a small market share. The $5.25 duty here would imply a sharp reduction in export volume, and the world price that appears most likely is four dollars per barrel. The same conclusion, that current prices cannot stand up, is drawn much more clearly in this case than in the base case.

TABLE 9
RESULTS OF HIGH SUPPLY ELASTICITY SIMULATIONS

ROYALTY AT PERSIAN GULF ($/BBL)	PERSIAN GULF AND NORTH AFRICA EXPORT VOLUME (MM BBL /DAY)	PERSIAN GULF AND NORTH AFRICA EXPORT REVENUE (BILLION $/YR)	OTHER OPEC EXPORT REVENUE (BILLION $/YR)
0 80	40 1	14 9	4 4
1 75	30 9	21 8	6 8
3 50	19 3	25 5	11 2
5 25	13 1	25 6	14 9
7 00	7 7	19 9	19 3
8 75	3 0	10 0	23 2

Tables 10 and 11 show the consumption, production, and cost in the regions of the world in this alternative case. A comparison with Tables 7a and 7b shows that only North American figures are affected, since the European and Japanese prices, and thus consumption, are set by the Persian Gulf duty. U.S. production in this case is two to three million barrels per day higher than in the base case.

TABLE 10

HIGH SUPPLY ELASTICITY CASE· CONSUMPTION AND PRODUCTION
(MM BBL./DAY)

A—CONSUMPTION			
DUTY/ REGION	UNITED STATES	EUROPE	JAPAN
0 80	24 4	19 9	6 8
1.75	22.0	17 5	5 8
3 50	19.1	15 1	4 7
5 25	17.6	13 8	4 0
7 00	16 4	12 8	3.6
8 75	15 3	12 2	3 4

B—PRODUCTION			
DUTY/ REGION	UNITED STATES	CANADA	EUROPE
0 80	7 8	1 2	4 4
1.75	9 5	1 5	4 4
3 50	12 3	2 0	4 4
5 25	13 9	2 1	4 4
7 00	15.5	2 4	4 4
8 75	17.3	2 7	4 4

At the $5.25 duty, for example, the U.S. crude price is $6.35, and production is 13.9 million barrels per day. Given the assumed Alaskan output of 2½ million barrels, this implies 11½ million from the lower 48, which compares to the current 9 million. These figures are thus still somewhat conservative, due to the 3-percent production decline assumed at unchanged prices.

TABLE 11

HIGH SUPPLY ELASTICITY CASE COST OF OIL IMPORTS
(BILLION $/YR)

DUTY	COST OF IMPORTS TO			TOTAL PAYMENTS TO OPEC
	UNITED STATES	EUROPE	JAPAN	
0 80	13 6	11 7	4 9	29 3
1 75	13.5	14 4	6 4	28 6
3 50	9 3	18 5	7 9	36 7
5 25	3.9	22.0	9 1	40 5
7 00	-2 8	24 7	10 0	39 2
8 75	-12 7	26.7	10 7	33 2

An interesting implication of these runs is the importance of marginal U.S. supplies in affecting the overall world oil market situation. They show, for example, that U.S. policies which affect domestic production levels would have important implications for the optimal behavior patterns of the producer cartel.

☐ **Alternate cases: higher demand levels.** The runs presented so far have predicted OPEC export volumes far below previous forecasts. Some runs were done with income elasticities of demand set at 1.5 rather than 1, to find the implications of stronger demand for products for oil market conditions. Table 12 shows the simulated outcome under these conditions. This rather extreme change in exogenous conditions flattens out the revenue as a function of export duty curve. Revenues are unchanged through the $3.50 to $5.25 range, then tail off somewhat as taxes rise further.

The implications of these results are naturally more favorable to oil exporters, and total OPEC proceeds approach 60 billion dollars per year. Even in this case, however, a seven dollar duty is probably unsustainable due to the low production levels implied. A tax at the higher end of the profit maximizing range, such as

TABLE 12

RESULTS OF HIGH INCOME ELASTICITY SIMULATIONS

ROYALTY AT PERSIAN GULF ($/BBL)	PERSIAN GULF AND NORTH AFRICA EXPORT VOLUME (MM BBL /DAY)	PERSIAN GULF AND NORTH AFRICA EXPORT REVENUE (BILLION $/YR)	OTHER OPEC EXPORT REVENUE (BILLION $/YR)
0 80	50 7	19 0	4 4
1 75	40 2	28.5	6 8
3 50	27 2	35 9	11 2
5 25	18.6	35 8	15 7
7 00	13 3	34 0	19 4
8 75	8 1	26 0	23 7

five dollars, becomes easier to maintain since exports would run at 19 million barrels per day. This then becomes more likely relative to a $3.50 tax.

Finally, some runs with extremely inelastic demand and supply curves were made. All demand elasticities were set at —0.25, and North American supply elasticity was set at 0.25. Since demand and supply curves display constant elasticity (as was demonstrated in Figure 1), these assumptions imply increasing Persian Gulf and North African revenues through a wide range of increased export duty levels. The results of these runs were that Persian Gulf and North African producers maximized their revenues at a 21 dollars per barrel duty, earning 80 billion dollars per year from oil sales. This result should not be unexpected, of course; they say that if there is essentially inelastic excess world demand for oil, its producers can charge whatever they please. These runs do show that even with the most unfavorable possible demand and supply parameters, total oil payments by importers are less than extrapolations of previous import trends at the high price. This is because of the new Alaskan and North Sea production, and because even very low elasticities imply moderate decreases in consumption as prices increase by a factor of ten.

☐ **Conclusions.** Using an economic model of the world oil market, we have found that the price hikes of crude which occurred in late 1973 are not likely to persist in the future. The laws of supply and demand imply behavior on the part of oil importing nations which make maintenance of these prices difficult for the largest producing region, the Persian Gulf and North Africa, due to the problem of allocating reduced production levels among the individual countries. The results of simulations from the model suggest that an export duty of half the current level, about $3.50, is most likely to occur in the long run. Changing the assumptions behind the simulations to make them reflect conditions more favorable to a high price implied that a value closer to $5.00 was easier to maintain, and thus more likely to occur.

It should be emphasized, however, that the oil model as an analytical tool does not stand or fall on these conclusions. The model itself can be adapted to reflect much different assumptions about supply and demand conditions, and will provide a consistent forecast of the world pattern of prices, consumption, and trade under those circumstances. The results given here reflect cost and demand conditions summarized in Sections 2, 3, and 4, and the accuracy of the forecasts depends on them.

In light of serious concern in developed countries in early 1974 about the long-run consequences of high-priced oil, various proposals were made whose rationale depended on an assumption of a continuing high price. The results presented here indicate that either investment plans or political concessions made on the basis of an expectation of seven dollar oil or a fear of twelve dollar oil would likely be regretted in the long run. In any case, a consistent forecast of oil market conditions, which integrates the decision maker's assumptions about exogenous conditions with the structure of the market, is necessary background information for an assessment of the costs and benefits of any policy. It is the task of applied economic theory, of which commodity models such as this are a part, to provide such information, and thus become a useful input into national policy making.

Appendix

Interpretation of the oil model equilibrium

■ The first section presented a mathematical statement of the equilibrium conditions of the world oil market model. It was stated there that these conditions could be interpreted as the first-order conditions of a quadratic programming problem. The contention will be demonstrated here, and its relevance for the actual numerical calculation of equilibrium in the oil market will be shown.

From the first section, we repeat the conditions necessary for a long-run equilibrium to exist in the world oil market.

$$\eta = Rd + r \tag{1a}$$

$$-(T'\eta) \leqslant k \tag{2}$$

$$d = Tt \tag{3}$$

$$(k + T'\eta)t' = 0 \tag{4}$$

$$t \geqslant 0$$

$$d_i \leqslant 0 \text{ if } i \text{ corresponds to a refined product}$$

$$d_i \geqslant 0 \text{ if } i \text{ corresponds to crude oil.}$$

Equation (1a) is just the net supply equation, which says that the consumption of products and the production of crude are determined by their respective prices. The economic constraint that no activity can make excess profits in the long run is represented in equation (2). Equation (3) is the material balance equation, and it guarantees that refining and trade activities are operated at a level consistent with meeting net supply. The further condition that any activity whose gross revenue fails to cover economic costs including normal profits will not be operated, is given by (4). This set of equations and inequalities completely characterizes long-run equilibrium in the oil market. Their relationship to a certain quadratic programming problem will be shown in this appendix.

☐ **The quadratic programming problem.** Following Boot,[53] we can write the general quadratic programming (QP) problem as

$$\max \quad a'x - \frac{1}{2}x'Bx$$

subject to

$$Cx \leqslant d$$

and

$$\begin{aligned} x_i &\leqslant 0 \text{ if } i \in N \\ x_i &\geqslant 0 \text{ if } i \in P. \end{aligned}$$

In this, x is an n-length vector of unknowns; N and P are a partition of the set of integers from 1 to n; a is an n-length vector and B is an $n \times n$ positive semidefinite matrix of parameters; and C is a $k \times n$ matrix and d is a k-length vector of constraint parameters. The Kuhn-Tucker necessary and sufficient conditions for x^* to be an optimal solution for the program are, first, that x^* be feasible

$$\begin{aligned} &x_i^* \leqslant 0 \text{ if } i \in N \\ &x_i^* \geqslant 0 \text{ if } i \in P \\ &Cx^* \leqslant d. \end{aligned}$$

Second, there must exist an $n + k$ length vector u of Lagrangian multipliers, partitioned as $u = (V'w')'$, with V of length n and w of length k. This vector of multipliers must satisfy

$$\begin{aligned} &V = Bx^* + C'w - a \\ &V_i \geqslant 0 \text{ if } i \in N \\ &V_i \leqslant 0 \text{ if } i \in P \\ &w \geqslant 0 \\ &V'x^* = 0 \\ &w'(d - Cx^*) = 0. \end{aligned}$$

☐ **The oil model in the programming framework.** The QP problem associated with the world oil market model is

$$\max \quad (-r' \vdots - k')\begin{pmatrix} d \\ t \end{pmatrix} - \frac{1}{2}(d'\ t')\begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix}\begin{pmatrix} d \\ t \end{pmatrix}$$

subject to

$$(-I \vdots T)\begin{pmatrix} d \\ t \end{pmatrix} \leqslant 0$$

$$\begin{aligned} &t \geqslant 0 \\ &d_i \leqslant 0 \text{ if } i \text{ corresponds to a refined product} \\ &d_i \geqslant 0 \text{ if } i \text{ corresponds to crude oil.} \end{aligned}$$

Using the Kuhn-Tucker conditions stated above, we characterize the solution of the program with the following conditions. The vector $(d^*\ t^*)$ corresponds to x^* in the general formulation of the QP problem, and the Lagrangian multiplier vector, $u = \begin{pmatrix} V \\ w \end{pmatrix}$ will be renamed $u = \begin{pmatrix} \begin{pmatrix} \delta \\ l \end{pmatrix} \\ \eta \end{pmatrix}$ so that $V = \begin{pmatrix} \delta \\ l \end{pmatrix}$ and $w = \eta$.

---

[53] In [8].

First, the optimal vector $(d^*\ t^*)$ must satisfy the constraints of the program:

$$d_i^* \leqslant 0 \text{ if } i \text{ corresponds to a refined product} \qquad \text{(A1)}$$

$$d_i^* \geqslant 0 \text{ if } i \text{ corresponds to crude oil}$$

$$t^* \geqslant 0$$

$$d^* \geqslant Tt. \qquad \text{(A2)}$$

Then, there are the conditions involving the Lagrangian multipliers:

$$\delta = Rd^* + r - \eta \qquad \text{(A3)}$$

$$d_i \geqslant 0 \text{ if } i \text{ corresponds to a refined product}$$

$$d_i \leqslant 0 \text{ if } i \text{ corresponds to crude oil}$$

$$l = T'\eta + k \geqslant 0 \qquad \text{(A4)}$$

$$\eta \geqslant 0 \qquad \text{(A5)}$$

$$\delta' d^* = 0 \qquad \text{(A6)}$$

$$l' t^* = 0 \qquad \text{(A7)}$$

$$\eta'(d^* - Tt^*) = 0. \qquad \text{(A8)}$$

We shall now proceed to give an economic interpretation to these conditions, and show how they correspond to the world oil model equilibrium conditions. Equation (A6), the complementary slackness condition for the net supply variables, states that if any product demand or crude supply is nonzero, its associated Lagrangian multiplier must be zero. The economics of the world oil market guarantees that the vector d will indeed be nonzero, so we conclude $\delta = 0$. Thus, condition (A3) can be restated $\eta = Rd^* + r$. We interpret η as the vector of commodity prices, so this condition simply says all net supplies must be on the net supply functions.

Similarly, we knew that any meaningful solution of world oil equilibrium will involve strictly positive prices. Thus, the complementary slackness condition (A8) implies $d^* = Tt^*$, or that neither crude nor products are thrown away. Finally, the vector of Lagrangian multipliers l is identified as the net loss associated with operation of each activity in the model. Conditions (A4) and (A7) have exact counterparts in the original statement of the oil model, for they show that no activity makes an excess profit, and that no activity that would make less than a normal profit is operated.

Incorporating the economic considerations just given, the Kuhn-Tucker conditions for the QP problem can be restated:

$$d^* = Tt^*$$

$$d^*_i \leqslant 0 \text{ if } i \text{ corresponds to a refined product}$$

$$d^*_i \geqslant 0 \text{ if } i \text{ corresponds to crude oil}$$

$$\eta = Rd^* + r$$

$$l = k + T'\eta \geqslant 0$$

$$l't^* = 0$$

$$t^* \geqslant 0.$$

These are exactly the conditions for long-run equilibrium in the world oil market as presented in Section 1.

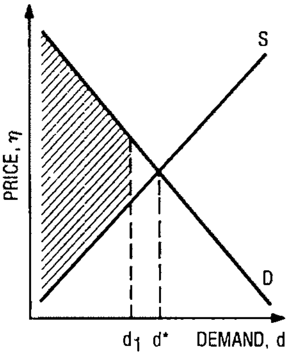
The criterion function for the QP associated with the equilibrium conditions in the world market can be restated as

$$-r'd - \frac{1}{2}d'Rd - k't.$$

This quantity, which Samuelson[54] identified as net social pay-off, is the sum of consumer's and producer's surplus less the full cost of refinery and transportation operation, and all taxes.

FIGURE A1

SUPPLY AND DEMAND FOR A ONE COMMODITY MARKET



The intuitive reasonableness of this criterion can be seen in a one-good model. Figure A1 shows the simple supply and demand diagram for a one commodity market with equilibrium at d^*.

The maximand of the associated QP in this case is the integral of the demand curve less the integral of supply curve, which is the shaded area in Figure A1. This quantity is obviously maximized at d^*, the competitive equilibrium, where it is equal to consumer's surplus plus producer's surplus.

Takayama and Judge[55] have an extended discussion of the expression of multicommodity equilibrium problems in the programming framework, with examples coming mostly from agricultural products. The reader is referred to them for more elaboration of this topic.

☐ **Solution of the quadratic program.** Since the equilibrium conditions of the oil market model can be interpreted as necessary and sufficient conditions for optimization of a certain QP, the large literature on QP solution was used to find an efficient method for determining oil market equilibrium. Reviews of this literature can be found in Boot and in Zangwill.[56] The method used here is related to the capacity method developed by Houthakker.[57] A detailed account of the method can be found in Kennedy.[58]

References

1. Adams, F. G. and Griffin, J. M. "An Econometric-Linear Programming Model of the U.S. Petroleum Refining Industry." *Journal of the*

---

[54] In [42].

[55] In [49].

[56] In [8] and [56], respectively.

[57] In [25].

[58] In [31].

American Statistical Association, Vol. 67, No. 339 (September 1972), pp. 542–551.

2. Adelman, M. A. "Is the Oil Shortage Real?" *Foreign Policy*, No. 9 (Winter 1972), pp. 69–107.
3. ———. *The World Petroleum Market*. Baltimore: Johns Hopkins Press, 1972.
4. Akins, J. E. "This Time the Wolf Is Here." *Foreign Affairs*, Vol. 51, No. 3 (April 1973).
5. Balestra, P. and Nerlove, M. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica*, Vol. 34, No. 3 (July 1966), pp. 585–612.
6. Baughman, M. L. *Dynamic Energy System Modelling—Interfuel Competition*. Washington, D.C.: National Science Foundation, 1972.
7. Bonner and Moore Associates. *U. S. Motor Gasoline Economics*. Washington, D. C.: American Petroleum Institute, 1967.
8. Boot, J. C. G. *Quadratic Programming*. Amsterdam: North Holland, 1964.
9. Bradley, P. G. *The Economics of Crude Petroleum Production*. Amsterdam: North Holland, 1967.
10. Chow, G. C. "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, Vol. 28 (July 1960), pp. 591–605.
11. Darmstader, J. "Appendix" in Schurr, S. H., ed., *Energy, Economic Growth and the Environment*, Baltimore: Johns Hopkins Press, 1972.
12. de Chazeau, M. G. and Kahn, A. E. *Integration and Competition in the Petroleum Industry* New Haven: Yale Univ. Press, 1959.
13. Dhrymes, P. J. *Distributed Lags: Problems of Estimation and Formulation*. San Francisco: Holden Day, 1971.
14. E. N. I. *Energy and Hydrocarbons (Energy Yearbook) Statistical Supplement*, 1972.
15. Erickson, E. W. and Spann, R. M. "Supply Response in a Regulated Industry: The Case of Natural Gas." *The Bell Journal of Economics and Management Science*, Vol. 2, No 1 (Spring 1971), pp. 94–121.
16. Fisher, F. M., Cootner, P. H. and Baily, M. N. "An Econometric Model of the World Copper Industry." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 568–609.
17. Goldberger, A. S. and Gamaletsos, T. "A Cross Country Comparison of Consumer Expenditure Patterns." Social Systems Research Institute, University of Wisconsin, November 1967.
18. Gordon, R. L. *The Evolution of Energy Policy in Western Europe: The Reluctant Retreat from Coal*. London: Praeger, 1970.
19. Grayson, C. J. *Decisions under Uncertainty: Drilling Decisions by Oil and Gas Operators*. Cambridge: Harvard Univ. Press, 1960.
20. Griffin, J. M. *Capacity Measurement in Petroleum Refining*. Lexington, Mass.: D C. Heath and Co., 1970.
21. Griliches, Z. "Distributed Lags: A Survey." *Econometrica*, Vol. 35, No. 1 (January 1967), pp 16–49.
22. ———. "Specification Bias in Estimates of Production Functions." *Journal of Farm Economics*, Vol. 39, No. 1 (February 1957), pp. 8–20.
23. Hartshorn, J. E. *Oil Companies and Governments*. London: Farber and Farber, 1962.
24. Hengstebeck, R. J. *Petroleum Processing*. New York: McGraw-Hill, 1959.
25. Houthakker, H. S. "The Capacity Method of Quadratic Programming." *Econometrica*, Vol. 28, No. 1 (January 1960), pp. 62–87.
26. ——— "New Evidence on Demand Elasticities." *Econometrica*, Vol. 33, No. 2 (April 1965), pp. 277–288.
27. ——— and Kennedy, M. "Demand for Energy as a Function of Price." Paper presented to the American Association for the Advancement of Science, February 1974.
28. Houthakker, H. S., Verleger, P. K., Jr., and Sheehan, D. P. "Dynamic Demand Analyses for Gasoline and Residential Electricity." Forthcoming in the *American Journal of Agricultural Economics*, 1974.
29. Jorgenson, D. W. "Anticipations and Investment Behavior," in Duesen-

berry, *et al.*, eds., *The Brookings Quarterly Model of the United States*, Amsterdam: North Holland, 1965.

30. Kaufman, G. M. *Statistical Decision and Related Techniques in Oil and Gas Exploration.* Englewood Cliffs, N. J.: Prentice-Hall, 1963.
31. Kennedy, M. "An Economic Model of the World Oil Market." Unpublished Ph.D. dissertation, Harvard University, 1974.
32. Mancke, R. B. "The Long-Run Supply Curve of Crude Oil Produced in the United States." *Antitrust Bulletin*, Vol. 15 (Winter 1970), pp. 277–756.
33. Manne, A. S. "A Linear Programming Model of the U.S. Petroleum Refining Industry." *Econometrica*, Vol. 26, No. 1 (January 1958), pp. 67–106.
34. ———. *Scheduling of Petroleum Refinery Operations.* Cambridge: Harvard Univ. Press, 1956.
35. Marshak, T. A. "A Spatial Model of U. S. Petroleum Refining," in A. S. Manne and H. M. Markowitz, *Studies in Process Analysis*, New York: Wiley, 1963.
36 McDonald, S. L. *Petroleum Conservation in the United States: An Economic Analysis.* Baltimore: Johns Hopkins Press, 1971.
37. Mikdashi, Z. et al *Continuity and Change in the World Oil Industry.* Beirut: Middle East Research and Publishing Center, 1970.
38. National Petroleum Council. *U. S. Energy Outlook.* Washington, D. C.: N.P.C., 1972.
39. Nelson, W. L. *Petroleum Refinery Engineering.* New York: McGraw-Hill, 1958.
40. Organization for Economic Cooperation and Development. *Energy Policy: Problems and Objectives.* Paris: OECD, 1966.
41. *Platt's Oil Price Handbook*, 1973.
42. Samuelson, P. A. "Spatial Price Equilibrium and Linear Programming." *The American Economic Review*, Vol. 42, No. 3 (June 1952), pp. 283–303.
43. Scherer, F. M. *Industrial Market Structure and Economic Performance.* Chicago: Rand McNally, 1970.
44. Schurr, S. H. et al. *Middle East Oil and the Western World.* New York: American-Elsevier, 1971.
45. Shell Oil Company. *The National Energy Outlook*, 1973.
46. Steele, H. Testimony in "Hearings on Government Intervention in the Market Mechanism: The Petroleum Industry," U. S. Senate, Judiciary Committee. Washington, D. C.: U. S. Government Printing Office, 1969.
47. Stocking, G. W. *Middle East Oil.* Kingsport: Vanderbilt Univ. Press, 1970.
48. Symonds, G. H. *Linear Programming—The Solution of Refinery Problems.* New York: Esso, 1955.
49. Takayama, T. and Judge, G. G. *Spatial and Temporal Price and Allocation Models.* Amsterdam: North Holland, 1971.
50. Tanzer, M. *The Political Economy of International Oil and the Underdeveloped Countries.* Boston: Beacon Press, 1970.
51. U. S. Bureau of Mines. *International Petroleum Annual 1972.* Washington, D. C.: U. S. Government Printing Office, 1974.
52. U. S. Bureau of Mines. *Minerals Yearbook.* Washington, D. C.: U. S. Government Printing Office, 1972.
53. U. S Cabinet Task Force on Oil Import Control. *The Oil Import Question.* Washington, D. C.: U. S. Government Printing Office, 1970.
54. U. S. Congress. "Investigation of the Petroleum Industry." U. S Senate, Committee on Government Operations. Washington, D. C.: U. S. Government Printing Office, 1973.
55. Working, E. J. "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics*, Vol. 41, No. 2 (February 1927), pp. 212–238.
56. Zangwill, W. I. *Nonlinear Programming.* Englewood Cliffs, N. J : Prentice-Hall, 1969.
57. Zannetos, Z. S. *The Theory of Oil Tankship Rates.* Cambridge: The M.I.T. Press, 1966.

Energy, environment, and economic growth

Anne P. Carter

Fred P. Pomerantz Professor of Economics
Brandeis University

A closed dynamic input-output model is used to appraise the effects of specific pollution abatement and new energy technologies on the rate of economic growth and on the relative importance of sectors in the U. S. economy over the next 10 to 15 years. Three sets of structural changes are considered: (1) expected changes in technology for electric power generation, transmission, and distribution; (2) reliance on coal gasification for additional fossil fuel consumption; and (3) pollution abatement to meet stricter standards for six types of air, water, and solid waste pollution. All of these changes tend to reduce the long-run growth potential of the system. Taken separately, none of them has a striking effect on the capital and output proportions of the economy nor on its growth rate, but their combined effect could be substantial. However, the computations show that the effects of increased pollution abatement and energy costs could be offset by modest reductions in the growth of energy consumption or by increased rates of consumer savings.

1. Introduction

■ Recent environmental problems and energy shortages have prompted questions about the future course of economic growth in the United States. Spokesmen for industry and government express concern lest capacity bottlenecks interfere with growth at the 3.5-percent rate that we have sustained over the past quarter century. Others challenge the desirability of the rapid economic growth that we have long taken for granted.

Economic growth is not a simple policy variable. Its future course will depend on many factors: on the availability of natural resources, now only vaguely appraised, both at home and abroad; on the expansion and composition of the labor force; on public and private consumption; on environmental policy; and on the specific techniques of production that will be used. While technology does not uniquely determine our development, it limits the economy's capacity to expand with any given consumption and resources. As

---

Anne P. Carter received the B.A. in economics from Queens College in 1945 and the Ph.D. in economics from Harvard-Radcliffe in 1949. Her current research centers on world resource problems and on income distribution.

This study draws on a series of studies undertaken at the Harvard Economic Research Project during 1971–1972 and continued at Brandeis and elsewhere in 1972–1973. These studies will be published in a forthcoming volume. The author wishes to thank Orani Dixon, Donald T. Jenkins, and Brian Dorsey for assistance at earlier stages and Carol T. Everett who was responsible for all of the computations at the final stage of analysis.

resource supplies and environmental standards change, new technologies are developed in response. In principle, these new technologies could increase or decrease the economy's growth potential. Actually, the "first generation" technologies addressed to today's environmental and resource conditions tend to decrease it. This study gives a rough quantitative appraisal of the implications of some specific pollution abatement and energy technologies that tend to reduce the rate of economic growth.

The data for this analysis are drawn from a series of studies that articulate individual abatement techniques and energy processes in a dynamic input-output framework.[1] Three sets of innovations are considered: projected changes in the technologies of generation, transmission, and distribution of electric power over the next ten years; coal gasification; a broad range of pollution abatement activities that would satisfy announced or projected standards for air, water, and solid waste pollution. While far from exhaustive, this is a fairly broad list of direct technological responses to environmental and resource conditions expected to come on line over the next decade. Their adoption could make important differences in our growth potential and in the relative importance of sectoral contributions to output and to the capital stock. If these changes are introduced, how will the economy of 1980–1985 differ from today's?

In a closed dynamic input-output system, the maximum rate of economic growth that can be sustained depends on all sectors' input structures, both on capital and on current account, including the level and composition of final consumption. One important conclusion of this study is that growth potential is more sensitive to changes in the propensity to consume than to increasing costs in the energy sectors. The expansion of the economic system is also constrained implicitly by the availability of labor and specific natural resources. Until very recently, the size of the labor force seemed to be the only major resource constraint that was binding on the U. S. economy. Since the work force was increasing at only one percent per year, continual increases in productivity were necessary to keep the effective labor supply sufficient for 3 to 4 percent growth. Thus labor-saving progress dominated technical change during the post-war period. The great bulk of industrial innovations—automation, computerization, material and design changes—represented direct or indirect economies of labor, with direct labor-saving predominating.[2] Had per capita consumption remained constant, increasing labor productivity might have freed capacity for increased capital accumulation and accelerated economic expansion. However, increasing productivity was offset by rising consumption per worker and thus the average rate of growth was fairly steady. Commoner[3] and others point out that much of the recent increase in labor productivity and in consumption has been achieved through more intensive exploitation of nonhuman resources that once seemed plentiful, particularly energy and the natural environment.

---

[1] See Carter [5].

[2] See Carter [4].

[3] See [6].

Saving labor at the expense of other natural resources may well have hastened the present "crisis" at a time when environmental and energy constraints have also begun to bind.

Just as labor-saving innovations forestall labor shortages, so our environment- and energy-oriented innovations address potential scarcities of natural gas, oil, clean air and water, etc. Precipitators, filters, and treatment plants are designed to protect the air and water; nuclear power and coal gasification are designed to supplement or supplant scarce supplies of clean fossil fuels; higher-voltage transmission and underground distribution lines should convey electric power to consumers with minimal environmental damage. While these new technologies, once installed, do not require much direct labor, they generally require more capital and intermediate inputs than the old technologies that they supersede or augment. These added requirements impose a drag on the growing system.

In sum, two types of influences tend to limit growth in the coming decade. First, increased requirements for delivering a given level of current consumption leave a smaller portion of productive capacity for capital formation. Second, environmental considerations and scarcities of nonhuman resources complicate the technological challenge of increasing the productivity of labor. Innovators must now deal with environmental problems as well as labor efficiency. It may well prove feasible to increase the efficiency of natural resources and labor simultaneously, but the problem is more difficult than that posed by labor alone. Unless effective labor and resource supplies increase in step with capital accumulation, plans to expand capacity cannot be realized.

To the extent that natural resource pressures actually reduce the rate of economic growth, there will be less pressure to increase labor productivity. If growth of industrial capacity were very slow, natural increases in the labor force alone might provide sufficient manpower for the growing stock of equipment. However, we shall show that minor changes in consumption could readily offset the increased intermediate and capital requirements of the new technologies so as to maintain or even increase the long-run growth rate. Thus, it is hard to predict whether the problems of environment and energy supply will tend to relax or to intensify historic pressures for ever-increasing labor productivity.

2. Computations

Analytical framework. A closed dynamic input-output model is the basis for all the computations:

$$(I-A)X - B\dot{X} = 0, \tag{1}$$

where X is a vector of total sectoral outputs and $\dot{X}$ is a vector of their time derivatives.

A and B are current account and capital coefficient matrices, respectively. The model is closed by the addition of a household row and column to A and B. The household row is a vector of coefficients representing income and indirect tax payments; the household column is a vector of coefficients representing expenditures by households and government plus net exports.

Equation (1) says that outputs are allocated to current account uses, AX, and to expansion, $B\dot{X}$. Since capital formation is endogenous, the sum of expenditures by the household industry is equal in national accounts terminology to gross national product minus gross private capital formation. If all sectors were to grow at a uniform rate, equation (1) could be rewritten:

$$(I - A - \lambda B)X = 0, \tag{2}$$

where λ is the uniform or "turnpike" growth rate. Tsukui[4] has shown that λ represents the unique uniform growth rate for the economy consistent with full capacity utilization of all sectors. This turnpike growth rate proves to be the maximum growth rate consistent with the given flow and capital coefficient matrices for a broad range of initial and terminal outputs. Corresponding to the turnpike growth rate, λ, is a set of output proportions, the positive eigenvector of the matrix $(I - A - \lambda B)$. Thus λ measures the economy's long-run growth potential with the specified input structures A and B, while the turnpike output proportions indicate rough norms for the relative importance of sectors on the uniform growth path. Tsukui and Brody[5] have estimated turnpike paths for Japan and the United States. Their findings suggest that both these economies operate reasonably close to the computed paths.

☐ **Computation procedures.** The iterative algorithm of Brody[6] was used to compute uniform growth rates and output proportions. We begin with a set of base year output proportions, X^0, estimate a trial value of λ and reestimate equilibrium output proportions, X^1, consistent with the first estimate of λ. In outline, the computation sequence is

$$\begin{aligned} \lambda^k &= \frac{(1, 1, \ldots, 1)(I - A)X^{k-1}}{(1, 1, \ldots, 1)BX^{k-1}} \\ X^k &= (A + \lambda^k B)X^{k-1}, \end{aligned} \tag{3}$$

where k is the number of the iteration. All computations converged within seven iterations.[7]

☐ **Comparative dynamics.** Structural changes in matrices A and B will of course bring changes in the turnpike growth rate and output proportions. In this study the economic impacts of structural changes are evaluated in terms of comparative dynamics. We establish benchmark growth rate and output proportions on the basis of capital and current account matrices representing the structure of production in the early 1970s. Successive changes in the base year matrices are introduced to simulate the adoption of new technologies of energy production and stricter standards of pollution abatement. The effects of these structural changes, singly and in combination, on λ measure their impacts on the economy's growth potential; changes in computed output proportions represent differential impacts of the specified structural changes on individual sectors.

---

[4] In [16].

[5] In [13] and [3], respectively.

[6] In [3].

[7] Computations were performed on the PDP-10 computer at the Feldberg Computer Center, Brandeis University, using the PASTIM matrix manipulation program of Richard Drost.

The effects of changes in intermediate input and capital coefficients differ depending on concomitant changes in the levels and patterns of final consumption. To illustrate the impact of changing consumption, alternate consumption vectors were introduced in combination with the changes in industrial coefficients. The alternative consumption patterns are discussed at the conclusion of Section 2.

☐ **Implementation.** In this portion of the paper we shall outline the base year matrices, the technological variants, and the consumption variants.

Base year matrices.

The 83-order 1970 coefficient matrix of the Interagency Growth Project[8] was chosen as the base year A matrix. This coefficient matrix is the result of a set of 1970 projections made in the middle sixties and is not based on actual 1970 statistics. The capital matrix is an update of the 1958 capital matrix of the Harvard Economic Research Project to 1970–1975 technologies by the Battelle Memorial Institute.[9] Updating incorporates engineering information and conjecture. The fact that the A and B matrices are both projections should not seriously impair their usefulness in this study, since they serve primarily as general points of reference for analyzing the impact of specific further changes. Both A and B are in 1958 prices. Annual replacement requirements were estimated and added to the A matrix. The current account and capital matrices were augmented to include a "household" row consisting of the sum of income payments and indirect taxes per dollar of output for each industry and a "household" column representing all purchases of final outputs (exclusive of gross capital formation) per dollar of net national income. The turnpike growth rate computed for the base A and B matrices was 3.5 percent per annum.

Technological variants.

The base matrices were modified by introducing coefficients that represent the new technologies. Quantitative estimates of new industrial structures are drawn from specialized studies by Istvan, Just, Jenkins, Berlinsky, Kok, and Dorsey.[10] Four major sets of structural changes were considered.

(i) The first consists of changes in technology for electric power generation, transmission and distribution. The input coefficients in the base year's flow coefficient matrix were replaced by coefficients for a projected 1980 input structure of the electric power sector. These new coefficients were estimated by Istvan.[11] The corresponding column of the base year capital coefficient matrix was similarly replaced by Istvan's estimates of 1980 capital coefficients. Given Federal Power Commission projections of future process mix, Istvan's projections show that increases in the relative importance of nuclear generation, additions of pollution abatement equipment, extra high voltage transmission and underground distribution lines will result in greater costs and, in particular, greater capital intensity in electric power delivery. He projects an increase in the total capital coefficient for the electric power industry from 3.2 to 4.8 and a 9-percent increase in real costs on current account (including replacement).

---

[8] See U. S. Department of Labor [17].

[9] See Fisher and Chilton [7].

[10] In [5].

[11] In [8].

(ii) The second set of modifications converts the increment in projected over present fossil-fuel consumption from electric utilities to gasified coal. If all additional consumption of fossil fuel took the form of gasified coal, then by 1985 the input coefficients of coal, petroleum, and natural gas into electric utilties would be cut to half of their former values. These coefficients were so cut and the coal gasification industry was then added to the base matrix. Its row has a single entry representing utilities' consumption of gasified coal equivalent in btu value to the reductions in conventional fossil-fuel consumption. The column represents Just's projected input structure for the Hygas process of coal gasification.[12] This represents the most speculative of all the new technologies included in this study. Because there is no on-line experience with Hygas or any of the other possible methods, the coefficient estimates are more speculative than estimates for, say, nuclear power generation, cooling towers or sanitary landfill of solid wastes. Environmental standards will require some process of coal gasification and/or liquifaction as reliance on coal increases. This process represents a feasible medium-cost variant of a number of coal gasification alternatives. In all probability, the volume of gasified coal assumed here is far greater than the amount we shall actually produce by 1985. To the extent that gasification capacity is lower, however, costs of alternative environmental protection policies would have to be taken into account.

(iii) Over the past quarter century electric power input coefficients into all consuming sectors have been increasing steadily at an average rate of roughly 3.5 percent per year[13] and Federal Power Commission projections of electricity consumption implicitly assume that these trends will continue over the next decades.[14] Increased electricity consumption is likely to intensify the impact of changing energy technologies. Computations representing the technological changes in (i) and (ii) were repeated with the base year energy row coefficients increased by 40 percent. The adjusted electric power row represents the economy's power consumption propensities a decade hence if present trends continue.

(iv) To simulate the impact of economy-wide pollution controls, the basic flow and capital coefficient matrices were augmented to include an abatement cost row and a dummy abatement column for each of six types of pollution: particulate air pollution, industrial water pollution, thermal pollution, municipal sewage disposal, strip mining pollution, and municipal solid waste disposal.[15] The levels of abatement activities are calculated to satisfy the abatement standards of the Clean Air Act of 1971 for particulates and requirements for primary treatment (elimination of 85 percent of biochemical oxygen demand and suspended solids) of municipal and industrial waste water.

---

[12] In [10].

[13] See Carter [4].

[14] See [18].

Consumption variants.

In the base matrix consumption proportions are those of 1970 and final consumption grows at the same rate as all other sectors on the turnpike path. Per capita consumption in the future will depend on the rate of growth of the population relative to the rate of growth of total final goods. In the absence of increases in the work force the economy can expand only as fast as increasing per capita productivity permits. Under these conditions, per capita real income increases at the same rate as the economy's uniform growth rate. Should population increase at the same rate as the economy, the rate of increase in per capita income would be zero. Over the next decade the rate of growth of the population will be positive but probably lower than the growth rate of the economy and rising per capita income is generally expected. Consumption patterns can be expected to change as per capita income rises and as product prices change. Because the household sector is large, the structure of consumption has a major influence on growth potential. The impact of changing industrial technologies was computed with each of four different consumption structures.

(i) The first is that of the base year.

(ii) The second makes a rough allowance for expected changes due to rising per capita income. Ten years hence, assuming 2.5-percent growth per annum in labor productivity, we can expect real income per capita to be 28 percent higher than it is now.[16] To estimate the effect of this rise in average per capita income on consumption patterns, we assume that spending patterns will remain the same as they are now for any given income level. But as average income goes up, the relative importance of high-income spending patterns will increase. To estimate one possible structure of future consumption we assume that all incomes will increase proportionally as GNP rises. Each income-specific spending pattern is then assigned a weight equal to the proportion of income received by those who earned 28 percent *less* income in the base year. The most striking difference between this projected structure of consumption and that of the base year lies in the personal savings rates. Because people with high incomes save much more than those in lower income groups, the percentage rate of savings under variant (ii) is triple that in the base year. History shows, however, that personal savings remained a remarkably stable percentage of the national income, rising only very slowly with time and/or per capita income. Consumers may be ready to save more in the future if environmental improvement requires it.

---

[15] See Leontief [12].

[16] Since the computations are illustrative, the exact increase in income per capita chosen here is somewhat arbitrary. Precise consistency with the computed growth rate could be achieved by an iterative process.

(iii) Here the savings rate is allowed to increase by only *one* percentage point over the base pattern (i) in keeping with historical trends. Proportions of final spending on goods and services are kept the same as in variant (ii).

(iv) Over the next ten years rising energy prices are likely to induce substitution in favor of less energy-intensive products. This variant superimposes energy-saving changes into the base year consumption patterns. Direct consumption of energy is cut by 20 percent while expenditures on other consumer items are increased proportionally so as to maintain the base year savings rate.

Table 1 lists the various scenarios that were computed.

TABLE 1

SCENARIOS COMPUTED IN THIS STUDY

CONSUMPTION STRUCTURE	(i) 1970		(ii) INCOME-SPECIFIC CONSUMPTION STRUCTURES REWEIGHTED FOR 28-PERCENT INCREASE IN REAL INCOME	(iii) ONE-PERCENT INCREASE IN SAVINGS OVER 1970	(iv) ENERGY-SAVING CHANGES
ELECTRICITY CONSUMPTION COEFFICIENTS	(a) 1970	(b) 1970 x 1.4			
TECHNOLOGY					
1 BASE YEAR	X	X	X	X	X
9 ALL POLLUTION ABATEMENT	X		X	X	X
10. 1980 ELECTRIC	X				
11 1980 ELECTRIC WITH COAL GASIFICATION	X	X			
12 ALL ABATEMENT, 1980 ELECTRIC AND COAL GASIFICATION	X	X	X	X	X

3. Results

■ **Effects of selected changes on λ.** Table 2 shows the effects on growth potential of specific technological shifts, separately and in combination, with the fixed base year consumption structure. Few of these changes affect growth potential by more than a few tenths of a percentage point. However, the cumulative effect of groups of changes is appreciable. Abating all water, air, and solid waste pollution to meet currently announced goals reduces λ from 3.5 to 3.0 percent. Projected changes in electric power generation and distribution technology are introduced separately and also combined with coal gasification. The combined changes lead to a 0.5 percentage point cut in λ. With pollution controls and changes in electric power generation combined, long-run growth potential is now reduced to 2.6 percent per year.

☐ **Effects of identified changes combined with rising industrial use of electric power.** The impact of electric power technology

TABLE 2

LONG TERM GROWTH POTENTIAL (λ) WITH NEW ENERGY AND ABATEMENT TECHNOLOGIES

STRUCTURAL CHANGE	λ PERCENT PER ANNUM
1. NONE (BASE YEAR STRUCTURE)	3.54
POLLUTION ABATEMENT	
2. PARTICULATE AIR POLLUTION (≃99 PERCENT)	3.44
3. INDUSTRIAL BOD AND SUSPENDED SOLIDS (PRIMARY TREATMENT)	3.47
4. MUNICIPAL BOD AND SUSPENDED SOLIDS (PRIMARY TREATMENT)	3.45
5. ACID MINE DRAINAGE	3.49
6. THERMAL POLLUTION (COOLING TOWERS)	3.52
7. ALL WATER POLLUTION	3.33
8. SOLID WASTE DISPOSAL	3.33
9. ALL POLLUTION	3.03
NEW ENERGY TECHNOLOGIES	
10. 1980 ELECTRIC POWER TECHNOLOGY	3.32
11. 10+ COAL GASIFICATION	3.06
12. ENERGY AND POLLUTION CONTROL (9+11)	2.59

SOURCE SEE TEXT

on the growth rate is magnified greatly when electric power usage expands. Table 3 shows how new technologies affect the growth rate when power consumption coefficients are increased. Istvan's projected changes in electric power technology alone reduce growth potential by two-tenths of a percentage point. If electric power consumption coefficients are multiplied by 1.4 in the base matrix, λ is reduced by a 0.6 percentage point. If this increase in electric power consumption is used in conjunction with Istvan's changes, λ is reduced by an 0.8 percentage point. The effects of increased power consumption are even more dramatic—a reduction of 1.0 percentage point—when the increase takes place in a technological scenario that includes coal gasification as well as Istvan's changes. The costs associated with projected increases in electric power usage are larger than those incurred from new developments in electric power technology. Hence, increasing power consumption cuts the growth rate more than the switch to new techniques. The tendency of these two developments to compound each other is striking. Separately, neither has an overwhelming impact on growth

TABLE 3

LONG TERM GROWTH POTENTIAL (λ) WITH VARYING ENERGY TECHNOLOGIES AND ELECTRICITY USE (PERCENT PER ANNUM)

ENERGY TECHNOLOGY	ELECTRICITY CONSUMPTION COEFFICIENTS	
	1970 (a)	1970 × 1.4 (b)
1. BASE YEAR	3.5	2.9
10. 1980 ELECTRIC	3.3	2.5
11. 1980 ELECTRIC WITH COAL GASIFICATION	3.1	2.1

SOURCE SEE TEXT

potential. The combined effect, 1.4 percentage points, is certainly large enough to be taken seriously.

☐ **Technological changes combined with changes in consumption.** "Households" is the largest sector of our model, comprising all public and private final consumption. At present its level of consumption is roughly five times the volume of gross private domestic investment. Hence, it is no surprise that growth potential is more sensitive to a given percentage change in household spending and saving patterns than to the same percentage change in individual industrial sectors. For each of the four consumption structures discussed in Section 2, Table 4 shows computed values of λ assuming all pollution controls, separately and in combination with new electric generating and coal gasification technology.

TABLE 4

LONG TERM GROWTH POTENTIAL (λ) WITH VARYING TECHNOLOGIES AND CONSUMPTION (PERCENT PER ANNUM)

TECHNOLOGY	CONSUMPTION STRUCTURE			
	(I)	(II)	(III)	(IV)
1. BASE YEAR	3.5	5.7	3.8	3.6
9. ALL POLLUTION ABATEMENT	3.0	5.1	3.3	3.1
12. ELECTRIC, COAL GASIFICATION AND ALL POLLUTION ABATEMENT	2.6	4.7	2.9	2.7

SOURCE: SEE TEXT

Table 4 confirms the sensitivity of the economy's growth to changes in consumption patterns. Reducing the energy content of final consumption (iv) greatly attenuates the impact of new energy technologies. Were household savings at any given real income level to remain constant as in variant (ii), increasing per capita income would raise savings to a point where 5- or 6-percent growth could be sustained, even in the face of our major "deteriorations" in technology. Greater affluence will mean spending more on entertainment and services, less on food and rent. The effect of such a shift in spending patterns tends to increase growth potential even when savings are restricted to only small percentage point increases over the base year [variant (iii)].

Figure 1 shows the tradeoffs between consumption and growth with and without pollution abatement and new energy technologies.

☐ **Output proportions.** By changing intermediate demand, new technologies alter the relative importance of individual sectors. Higher capital requirements of new energy and pollution control technologies might be expected to increase the relative importance of capital producing sectors. On the other hand, deceleration of economic growth means a smaller proportion of economic resources devoted to capital formation. Because these tendencies offset each other, it is difficult to make *a priori* judgments about the effects of the new technologies on the relative importance of individual sectors. Our computations give some indication of what they may be.

Figures 2 and 3 show the proportionate contributions of the new construction and iron and steel sectors to total gross output

FIGURE 1

TRADEOFFS BETWEEN CONSUMPTION AND GROWTH WITH CHANGING TECHNOLOGIES

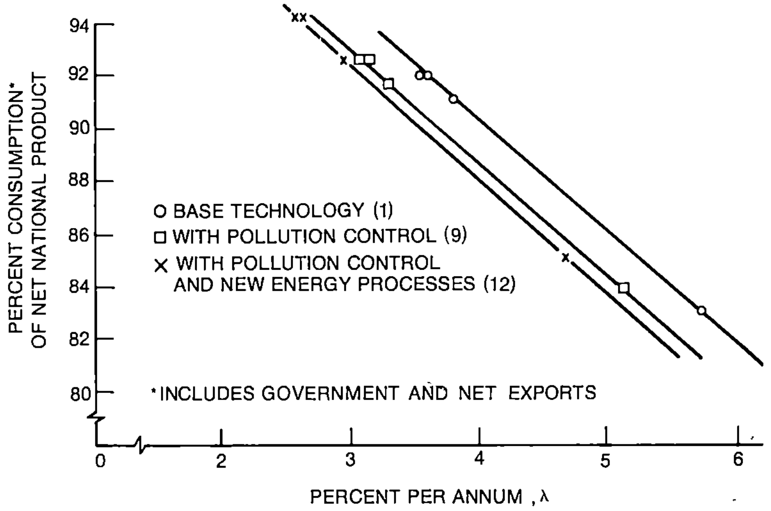


FIGURE 2

IMPORTANCE OF CONSTRUCTION WITH CHANGING GROWTH RATES

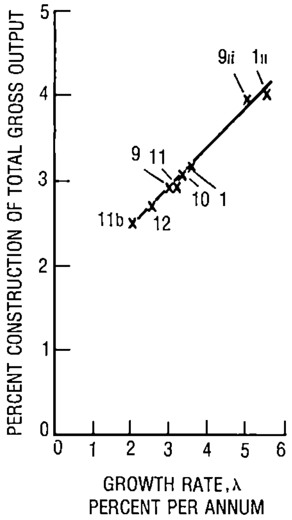
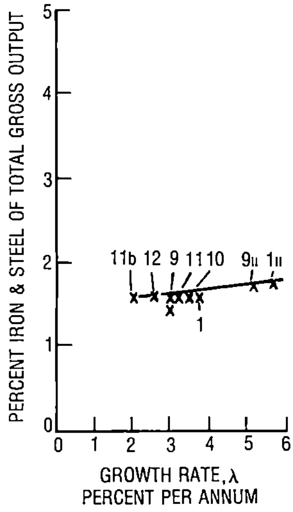


FIGURE 3

IMPORTANCE OF IRON AND STEEL WITH CHANGING GROWTH RATES



under various scenarios that we have computed. The proportion of total gross domestic activity is measured on the vertical axis while λ is shown on the horizontal axis. The circled number next to each point indicates a combination of technological and consumption structures identified in Tables 2, 3, and 4.

Table 5 shows that the relative importance of new construction in the economy varies from 2.5 to 4.1 for different assumptions. Since new construction is the largest single industrial component of most sectors' capital stocks and also of household capital, it is not surprising that this sector's importance varies directly with the growth rate. However, the relative importance of the iron and steel sector is constant, because steel is used in many consumer products (automobiles, cans, household equipment) as well as in capital goods.

TABLE 5

SECTORAL SHARES OF TOTAL GROSS OUTPUT WITH VARIOUS TECHNOLOGIES AND CONSUMPTION PATTERNS (PERCENT OF TOTAL GROSS OUTPUT)

	SCENARIO NUMBER*							
	11b	12	9	11	10	1	9(ii)	1(ii)
λ	2.1	2.6	3.0	3.1	3.3	3.5	5.1	5.7
PRODUCING SECTOR								
NEW CONSTRUCTION	2.5	2.7	2.9	2.9	3.0	3.1	3.9	4.1
MAINTENANCE CONSTRUCTION	2.2	2.1	4.4	2.1	2.1	2.2	2.1	2.1
FOOD	4.4	4.3	4.4	4.4	4.3	4.4	3.8	3.8
IRON & STEEL	1.6	1.6	1.6	1.6	1.6	1.6	1.8	1.8
AUTOMOBILES	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1
ELECTRIC UTILITIES	2.8	2.2	2.0	2.1	2.2	2.0	1.8	1.8

*SCENARIOS ARE LISTED IN TABLE 1

SOURCE SEE TEXT

The relative importance of most sectors is, like that of steel, virtually invariant with respect to the changes considered in this study. This is demonstrated in Table 5. Utilities increase and decrease in relative importance as electricity input coefficients are varied. Food declines in relative importance as high-income consumption patterns are introduced and the growth rate rises into the neighborhood of 5 percent. By and large, however, the proportions cited for each sector in Table 5 are remarkably stable. This tendency is explained by the limited range of the growth rates covered and the modest proportion of total resources devoted to growth under any of the options considered.

☐ **Changing composition of capital stock.** While the composition of annual output remains relatively stable with respect to our changes, a growing proportion of the capital stock is devoted to energy and abatement capital and the average capital-output ratio for the economy rises by as much as 6 percent over the base year economy. Table 6 shows that by and large higher capital intensity is characteristic of the lower growth rate scenarios. It also gives the proportion of total capital invested in utilities and in pollution abatement under various sets of assumptions. Utilities accounted for 9.9 percent of total investment in the base year. Initially they have a much greater capital-output ratio than most other sectors. Increasing investment requirements for utilities and increasing use of electric power add significantly to overall capital requirements.

TABLE 6

ECONOMY-WIDE CAPITAL/OUTPUT RATIOS AND PROPORTION OF INVESTMENT IN UTILITIES AND ABATEMENT WITH VARYING STRUCTURES (DOLLARS PER DOLLARS OF TOTAL GROSS OUTPUT)

SCENARIO NUMBER	GROWTH RATE (λ) (% PER ANNUM)	CAPITAL/ OUTPUT RATIO	PERCENT OF TOTAL INVESTMENT IN	
			UTILITIES	POLLUTION ABATEMENT
11b	2.1	0.84	15.3	*
1b	2.9	0.81	12.1	*
9_{I}	3.0	0.80	9.8	3.2
11a	3.1	0.81	11.8	*
1	3.5	0.79	9.9	*
12_{II}	4.7	0.82	10.8	2.5
9_{II}	5.1	0.80	8.9	3.2
1_{II}	5.7	0.78	9.1	*

*POLLUTION ABATEMENT COSTS ARE NOT ESTIMATED FOR THESE SCENARIOS
SOURCE SEE TEXT

4. Conclusions

■ The above results reveal the sensitivity of the economy—of its sectoral proportions and growth potential—to prospective structural changes. Taken separately, no one of the technological changes considered in this study exerts a decisive effect on the capital and output proportions of the economy nor on its growth potential. Nevertheless, their combined effect could be substantial. Furthermore, we have dealt with some of the more important but not all of the major adjustments that environmental and resource pressures will induce over the next decades. As the economy gets larger, progressively stricter environmental safeguards will be necessary simply to maintain any given standards of air and water quality. More substances will be recognized as pollutants. Most known abatement technologies involve very sharply increasing costs when the percentage of residual pollution is to be lowered. Significantly higher costs of extraction and refining of all fuels and ores are generally anticipated.

On the other hand, adaptive responses to rising energy and pollution costs soften the impact of these resource and environmental constraints. Our computations show, for example, that a 3.5-percent per year increase in electricity consumption coefficients greatly enhances the growth-limiting effects of new energy technologies. Conversely, a modest reduction in electricity coefficients could neutralize their impact. Ayres and Gutmanis[17] point out that current best practice technologies in many polluting sectors already show substantially lower gross emissions than average technologies now in use. The cumulative impact of small industrial and consumer adaptations to changing energy and environmental cost conditions is hard to anticipate. It could be very large.

Within the framework of this analysis we could study changing economic options as additional technological changes evolve. While the relative contributions of individual sectors do not change markedly in the face of the changes just considered, it would be unrealistic to ignore the lags and potential bottlenecks likely to arise in a sequence of major technological transitions. Leontief and Istvan[18] offer a more flexible framework for tracing year-to-year changes in outputs and investment that depart from the turnpike path.

The value of a general equilibrium approach extends beyond the limited area of technological policy to broader economic questions concerning the changing balance of income and tax policy, public spending, prices, and labor productivity. For some years, at least, the economy's growth potential with any given consumption structure will be lower than it used to be. However, this study has shown that a small once-for-all increase in consumer savings or other shifts in private or public final spending patterns can easily offset the drag. Thus, there remains a wide range of tradeoffs between faster growth and consumption at any given point. Cross-sectional evidence shows that high-income consumers save a substantially greater portion of their incomes than low. Were consumer savings to rise with income over time at even half the rate that they do in cross section, the growth rates of the past 25 years could easily be sustained. Actual future growth will depend on the complex policies that determine real savings for any given technological context.

Fundamentally, the turnpike growth rate measures the rate at which the economy's productive capacity—its capital stock—can be increased. Real growth also requires that the effective labor supply be sufficient to man the growing productive capacity. Increases in labor productivity have contributed more than growth in the labor force to rising output in recent years.[19] Recent studies suggest, however, that the rate of increase of productivity is decelerating.[20] Some of this deceleration may reflect resource pressures. Whatever its explanation, lagging productivity change could also limit growth. Here, as in the case of consumer savings, the economy retains a broad range of options. While both population and productivity increases are decelerating, rising labor force participation can easily offset their effect on growth potential. At a time when women are clamoring for a share in the job market, when youth suffer high rates of unemployment, and when firms experiment with the four-day week, labor shortages need not obstruct growth—not yet anyway.

---

[17] In [2].

[18] In [11] and [9], respectively.

[19] See Solow [15].

[20] See Nordhaus [14] and Almon *et al.* [1].

Pollution abatement might be viewed as a public good, i.e., as a form of consumption, but our conclusions do not rest on this assumption. We can meet environmental standards and resource constraints over the next decade and still maintain or even increase present growth rates for conventional goods. This should not be construed as advocacy of rapid economic expansion. The social and environmental evils of bigness are not all captured and dealt with in pollution abatement vectors; nor is it possible to represent all the tradeoffs between the present and the remote future, at home and abroad, in a growth model. Despite present constraints, we can still grow. Whether or not we should or want to is another question.

References

1. Almon, Jr., C., Buckler, M. B., Horwitz, L. M., and Reimbold, T. C. *1985 Interindustry Forecasts of the American Economy,* INFORUM Research Report No. 9. College Park, Md.: University of Maryland, Bureau of Business and Economic Research, August 1973 (preliminary)
2. Ayres, R. U. and Gutmanis, I. "Technological Change, Pollution and Treatment Cost Coefficients in Input-Output Analysis," in Ronald G. Ridker, ed., *Population, Resources, and the Environment,* vol. 3, The Commission on Population Growth and the American Future *Research Reports.* Washington, D. C.: U. S. Government Printing Office, 1972.
3. Brody, A. *Proportions, Prices and Planning.* Amsterdam: North-Holland Publishing Company, 1970.
4. Carter, A. P. *Structural Change in the American Economy.* Cambridge: Harvard Univ. Press, 1970.
5. ———, ed. *Structural Interdependence, Energy and the Environment,* forthcoming.
6. Commoner, B., Corr, M., and Stamler, P. J. "The Causes of Pollution." *Environment,* Vol. 13, No. 3 (April 1971).
7. Fisher, W. H. and Chilton, C *An Ex Ante Capital Matrix for the United States, 1970–1975.* Columbus: Battelle Memorial Institute, March 1971.
8. Istvan, R. *1890 Inputs for Private Electric Utilities.* Report to the Interagency Growth Project. Harvard Economic Research Project, Cambridge, August 1972, mimeographed.
9. ———. "Interindustry Impacts of Projected Electric Utility Capital Formation," in A. P. Carter ed., *Structural Interdependence, Energy and the Environment,* forthcoming.
10. Just, J. L. "Impacts of New Energy Technology Using Generalized Input-Output Analysis." Unpublished Ph.D. dissertation. Cambridge: Massachusetts Institute of Technology, 1972.
11. Leontief, W. W. "The Dynamic Inverse," in A. P. Carter and A. Brody eds., *Contributions to Input-Output Analysis,* Vol. 1, pp. 17–46, Amsterdam: North-Holland Publishing Company, 1970.
12. ———. "Environmental Repercussions and the Economic Structure: An

Input-Output Approach." *Review of Economics and Statistics,* Vol. 52, No. 3 (August 1970), pp. 262–271.

13. MURAKAMI, Y., TOKOYAMA, K., AND TSUKUI, J. "Efficient Paths of Accumulation and the Turnpike of the Japanese Economy," in A. P. Carter and A. Brody, eds., *Contributions to Input-Output Analysis,* Vol. 2, pp. 24–47, Amsterdam: North-Holland Publishing Company, 1970.
14. NORDHAUS, W. D. "The Recent Productivity Slowdown." *Brookings Papers on Economic Activity,* Vol. 3, 1972.
15. SOLOW, R. M. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics,* Vol. 39, No. 3 (August 1957), pp. 312–320.
16. TSUKUI, J. "Turnpike Theorem in a Generalized Dynamic Input-Output System." *Econometrica,* Vol. 34, No. 2 (April 1966), pp. 396–407.
17. U. S. DEPARTMENT OF LABOR, BUREAU OF LABOR STATISTICS. *Projections 1970—Interindustry Relationships—Potential Demand Employment,* Bulletin No. 1536. Washington, D. C.: U. S. Government Printing Office, 1966.
18. U. S. FEDERAL POWER COMMISSION, BUREAN OF POWER. *Trends and Growth Projections of the Electric Power Industry.* Washington, D. C.: U. S. Government Printing Office, 1969.

SHORT ARTICLES, COMMENTS AND REVIEWS

595 Railroad diversification: where lies the public interest?
G. C. EADS

614 The losses on savings deposits from interest rate regulation
D. H. PYLE

623 The implications of regulation for induced technical change
V. K. SMITH

633 The regulatory implications of three alternative econometric supply models of natural gas
R. S. PINDYCK

646 Determinants of research and development activity by electric utilities
R. P. WILDER and S. R. STANSELL

651 Dynamic aspects of public goods: a further analysis of the telephone system
B. von RABENAU and K. STAHL

670 Insurance as a market in contingent claims: structure and performance
J. M. MARSHALL

683 The economics of product safety and imperfect information
V. P. GOLDBERG

689 The economics of product safety: a rejoinder
W. Y. OI

696 On "duration" and the optimal maturity structure of the balance sheet
M. A. GROVE

710 The behavior of a firm subject to stochastic regulatory review: comment
W. T. ZIEMBA

713 The behavior of a firm subject to stochastic regulatory review: correction
A. K. KLEVORICK

715 Lieb's *Freight Transportation: A Study of Federal Intermodal Ownership Policy*
H. A. HOROWITZ

Railroad diversification: where lies the public interest?

George C. Eads

Associate Professor of Economics
The George Washington University

The potential public benefits and costs of railroad diversification are examined in this article. The principal claimed benefit, that diversification will produce an increased flow of capital to the railroad industry, is unlikely to be realized. The principal cost, that management may use diversification as a means of transferring assets from railroading to other more profitable lines of business, is less a problem created by diversification than a reflection of the current financial state of the railroad industry. Railroad diversification should be neither banned nor tightly regulated. Either will further hamper a much needed process of adjustment that already has been too long delayed. Instead, diversification ought to be welcomed, for it will place a needed check on the ICC's ability to force railroads to operate in a noneconomic fashion.

1. Introduction

■ Public concern over the possible consequences of diversification by railroads is not new. In 1906 Congress passed the Hepburn Act which contained the so-called "commodities clause" declaring it unlawful:

> . . . for any railroad to transport [in interstate or foreign commerce] any article or commodity other than timber and the manufactured products thereof manufactured, mined, or produced by it or under its authority, or which it may own in whole or in part or in which it may have an interest, direct, or indirect, except such articles or commodities as may be necessary for its use in the conduct of its business as a common carrier.[1]

This clause was enacted because of fears that the railroads, having at that time a virtual monopoly over inexpensive transportation in some sections of the country, might utilize this power to favor the products produced by railroad-controlled firms, thereby extending their monopoly power.[2] The clause severely restricted the ability of the railroads to diversify outside of the field of transportation and rendered the railroads unattractive as take-over targets by nontransportation firms.

In the early 1960s, the "commodities clause" ceased to be an effective bar to railroad diversification when the ICC declared that it lacked jurisdiction over railroad holding companies since they

---

The research leading to this paper was made possible in part by an NSF Institutional Grant. An early version was presented at the Western Economics Association meetings in August 1973. I would like to thank James Barth, Robert Goldfarb, George Hilton, and Robert Tollison for their helpful comments and assistance. I am responsible for any mistakes that remain.

[1] 49 U.S.C.A., Sec. 1(8).

were not carriers.[3] The railroads rushed to take advantage of this loophole. Between 1962 and the end of 1969, railroads whose assets comprised almost half the assets of all Class I railroads[4] had formed holding companies for the purpose of engaging in diversification.[5] In some cases, diversification activities have been limited in scope, and railroading has remained the principal line of business.[6] In others, however, diversification has proceeded to the point where the nonrailroad portion of the business generates a substantial majority of total group operating revenues and an even higher percentage of group profits.[7] In two cases the parent holding company has seen fit to sell off the railroad to third parties, thereby diversifying entirely out of the railroad business.[8]

However, regardless of whether it should be considered as representative, it was the diversification activity of the Pennsylvania Railroad, both before and after its 1968 merger with the New York Central, that attracted the greatest public attention. After the Penn Central declared bankruptcy in June of 1970, it was charged that part, if not all, of the railroad's problems could be traced to a drain of cash and management talent caused by the carrier's extensive diversification program. The staff of the House Banking and Currency Subcommittee went so far as to title one section of its report on the Penn Central failure "Diversification: Road to Ruin."[9]

The concern that the Penn Central's diversification activities might have contributed to its bankruptcy cast a shadow of suspicion over the diversification activities of all regulated enterprises. The staff of the ICC, after having investigated the diversification activi-

---

[2] George Hilton has pointed out to me that the precise impetus was the fear of the consequences of the control that several railroads then had over large deposits of anthracite coal.

[3] "ICC Clears Way for Kansas City Southern to Diversify Outside Transportation Field," *The Wall Street Journal*, April 23, 1962, p. 22.

[4] A Class I railroad is a railroad with annual revenues of $5 million or more. Class I line-haul railroads haul about 99 percent of all line-haul rail traffic, operate 94 percent of the mileage, and employ about 92 percent of the railroad personnel. See [18], p. a5.

[5] In ICC [11], Table 1, p. 11.

[6] In 1971, 87 percent of the revenues of Santa Fe Industries and 84 percent of the revenues of the Southern Pacific Company came from their respective railroads. See Direct Exhibits of the Department of Transportation in the Air Carrier Reorganization Investigation, Docket 24283 *et al.*, Exhibit DOT-DE-3, pp. 48 and 61.

[7] In 1971, IC Industries received 46 percent of its revenues and 41 percent of its profits from the Illinois Central Railroad; the Missouri-Kansas-Texas Railroad contributed 43 percent of Katy Industries' revenues. In 1968 the Bangor and Aroostook Railroad contributed only six percent of the revenues and one percent of the profits of its parent holding company, Bangor Punta. See DOT-DE-3, pp. 17 and 27, and "Special Review of Railroad Conglomerates," Report from M. Paolo, Director, Bureau of Accounts, ICC, to Chairman Brown and Vice Chairman Stafford, dated March 11, 1969, pp. 34–35.

[8] In 1969 Bangor Punta sold the Bangor and Aroostook Railroad to the Amoskeag Company. In 1972 Northwest Industries completed the sale of the Chicago and Northwestern Railroad to Northwest Employees Transportation Company (NETCO), an employee-owned firm.

[9] See [24], p. 23. The staff concluded that the Penn Central's diversification program ". . . resulted in a net cash drain from the railroad of approximately $175 million." See also Daughen and Binzen [5], pp. 214–224.

ties of regulated motor carriers, stated that it had uncovered "several potential problem areas" and recommended ". . . strengthening of the Commission's power to obtain information concerning, and exert control over, transactions including potential abuses."[10] The Civil Aeronautics Board, citing the "host of questions raised by the Penn Central collapse," instituted a broad-ranging investigation of the effects of air carrier diversification on "the Board's ability to discharge effectively its regulatory responsibilities."[11] And in March of 1971, the Executive Committee of the National Association of Regulatory Utility Commissioners (NARUC) appointed an Ad Hoc Committee on Nonutility Investments ". . . to examine the trend toward diversification among regulated companies and to determine the effect of such diversification on public utility regulation."[12]

We have two reasons, therefore, for focusing on railroad diversification in this paper. First, because it has been blamed in part for the railroad industry's current financial plight. Second, because the experiences of the railroads may shed light on the issues that are likely to emerge as firms in other regulated industries expand the scope of their business outside traditional areas. Before intelligent public policy concerning diversification by regulated industries can be formulated, the factors inducing the desire to diversify, the potential benefits of diversification to the public at large, and the dangers to the public interest alleged to result from diversification all need to be spelled out and critically examined. This is done for the railroads in the following three sections of this paper. In Section 5 we review the relatively meager evidence that exists concerning the actual effect that diversification has had upon railroad performance. Finally, we present our conclusions on whether railroads should be permitted to diversify and outline some of the lessons the railroad experience seems to offer for other regulated industries.

2. Reasons for carrier interest in diversification

■ The principal motivating factor behind railroad interest in diversification is not difficult to ascertain. Between 1906, when diversification was barred by the enactment of the "commodities clause," and the 1960s, the monopoly power of the railroads was severely eroded by the development of alternative modes of transportation, principally trucking.[13] However, both the ICC and the railroad unions prevented the industry from adjusting fully to this changed competitive situation. The railroads were obliged to continue to provide service to many points where traffic no longer warranted it and were required to employ workers whose usefulness was questionable. In the Northeast, where the development of new competition and shifts in the location of industry hit the railroads particularly hard, some of the carriers plunged into bankruptcy. Given such a picture, it is not surprising that a Department of Transportation (DOT) study of five railroad holding companies concluded that ". . . [t]he common denominator of railroad holding company formation seems to have been the desire to reduce shareholder vulnerability to the low rates of return, and the cyclicality of earnings which have become typical of the railroad industry."[14]

---

[10] See ICC [11], p. 25.

[11] See Civil Aeronautics Board Order 72-3-27, March 10, 1972, p. 4.

[12] See [20], p. 2.

[13] See Friedlaender [7] and Meyer *et al.* [16]. Moore is willing to go further than this. He states: "The monopoly power that the railroads had in short-haul markets during the nineteenth century no longer exists because of truck competition. Any monopoly power remaining with the railroads is limited to long-haul, bulk-commodity traffic. But such traffic by its very nature is inherently competitive. For example, a grain producer in the

However, interest in diversification has not been limited to those railroads showing the worst earnings records. Certain of the relatively prosperous carriers, such as the Union Pacific, the Southern Pacific, and the Chessie System also have formed holding companies as vehicles for diversification. And as has been noted above, firms in the regulated motor carrier and airline industries also have engaged in diversification even though neither of these industries has been plagued by the disastrous earnings record that has haunted the railroad industry throughout most of the postwar period. Consequently, while the reasons listed by the DOT doubtless are the primary ones stimulating railroad interest in diversification, others certainly have been present.

A 1969 ICC staff study cited several reasons for motor carrier interest in diversification which reduce to a perceived need to lower vulnerability to business downturns and improve access to capital markets by reducing dependence on a single line of business. Additional factors mentioned were the desire to attract better management than might be drawn to a firm specializing in a single line of business and a desire to achieve certain economies realizable through expansion into closely related lines of business.[15]

The stated reasons for air carrier interest in diversification parallel those of the motor carriers. Earnings in the airline industry have displayed a particularly high degree of cyclicality. The carriers have argued that diversification into lines of business displaying either more stable, or preferably, counter-cyclical earnings patterns would reduce their cost of capital. They, too, mention diversification as a means of attracting and holding higher quality management talent and point to certain economies that can be achieved through diversification into functionally related lines of business.[16] An additional factor stimulating airline interest in diversification, and one which has also been important to the railroads, is the desire to use

---

Middle West has the choice of sending his grain to the Gulf Coast, to the East Coast, or to the Great Lakes for export, or to numerous other regions for domestic consumption. As a consequence, even though only one road might exist between the shipper and a particular destination, the competition of other roads and other destinations is always potent. . . . With motor carriers ready to offer transportation even for long hauls if rail rates get too high, railroads have very limited power over the rates of bulk commodities," [19], pp. 91–92.

[14] Testimony of Russell F. Murphy, Chief, Industry Analysis Division, Office of Policy and Plans Development, United States Department of Transportation in The Air Carrier Reorganization Investigation, Docket 24283 *et al.*, November 24, 1972. Exhibit DOT-T-1, p. 3.

[15] See [11], p. 30, and [10], pp. 13–19.

[16] See, for example, Testimony of A. M. deVoursney, Senior Vice President, Corporate Planning, United Airlines, in the Air Carrier Reorganization Investigation, Docket 24283, Exhibit U-T-5, pp. 11–12.

substantial tax benefits consisting of investment tax credits and loss carryforwards.[17]

The management of each regulated transportation enterprise denies that its interest in diversification signals a decline in its interest in the regulated portion of its business. Railroad management has been particularly vocal on this point. Citing the poor recent earnings record of most railroads, the management has argued that only by diversifying into more profitable lines of business can the industry rebuild its credit position and thereby gain access to the capital is requires for modernizing its plant and equipment.[18]

3. Potential benefits to the public from diversification

■ A successful program of diversification undoubtedly may be of substantial benefit to the stockholders of a regulated enterprise; however, neither the customers of the enterprise nor the general public necessarily also benefits. Indeed, attempts to isolate non-stockholder benefits have not met with much success.

One of the principal benefits claimed for diversification is that it may lead to reduced capital costs.[19] A reduction in capital costs to the railroads either would permit a reduction in rates or, more importantly, would make it easier for the railroads to raise the large sums of capital they require to modernize their operations and improve service. Improvements in either of these two areas obviously would qualify as public benefits.

In view of the frequency with which this claim is heard and the importance placed upon it by regulators, it is necessary to determine just how any such reduction might come about as a result of diversification. Two possible avenues have been suggested. If the capital markets discriminated inappropriately against regulated firms, and if diversification provided a means by which this "imperfection" might be overcome, the cost of capital to these firms might be lowered as a result.

Consider, for example, a case in which restrictions contained in a carrier's credit agreements prevent it from borrowing additional funds even though the carrier's management believes that profitable uses of funds exist. This might reflect either a market imperfection or merely a difference in risk preference between the carrier's bondholders, anxious to protect the security of their original investment, and the carrier's management. Nevertheless, if the carrier were part of a diversified firm, and if the carrier's management could convince the top management of the parent that the proposed investment was justified, it might be possible to overcome the credit restrictions either by means of a direct loan from the parent or another subsidiary or by a pledge of the assets from one of these sources to back up additional borrowing by the carrier. Either way, the carrier's potential borrowing power would be expanded. It is unrealistic, however, for regulatory agencies to expect any such advantages to flow only in one direction. If another subsidiary of the parent firm could present a convincing case for allocating additional funds to it even though restrictions in its credit agreements limited its borrowing power, top management could be expected to draw upon the carrier's resources to support this other subsidiary. This might involve loans from the carrier to the other subsidiary or pledges of the carrier's assets to support additional borrowings, provided such pledges are possible under the terms of the carrier's own credit agreements. Thus, the carrier might be exposed to greater risk, balanced by the prospect of greater profits for the firm as a whole.

---

[17] As of June 30, 1972, total unused investment tax credits available to the major airlines amounted to $245 million. [Statement of Russell F. Murphy, Exhibit DOT-T-1, p. 25, note 2.] TWA alone reported that it had "about $225 million [consisting of $61 million in unused accumulated investment tax credits and $165 million in tax-loss carryforwards] it could potentially use in an acquisition program." See "TWA Tax Credits Figure," *The Wall Street Journal*, Friday, November 10, 1972, p. 31. In June of 1973, TWA announced that it had agreed to purchase Canteen Corporation from ITT for a price of $132 million. See *The Wall Street Journal*, June 7, 1973, p. 15.

[18] See, for example, "Letter to Stockholders," in Penn Central Company's Annual Report, 1969, p. 3.

[19] See, for example, "Diversification as a Means of Financial Support for Surface Transportation Utilities," Remarks by Kenneth H. Tuggle, Member of the Interstate Commerce Commission before the Section of Public Utility Law and Section of Administrative Law, American Bar Association Annual Meeting, San Francisco Hilton Hotel, Tuesday, August 15, 1972, pp. 11–12.

Alternatively, suppose that poor recent performance by the carrier's industry led potential investors to attach more stringent conditions to the use of their funds than might be warranted solely on the basis of the carrier's own performance. If the carrier were a division of a diversified firm, this "industry effect" might be short-circuited and capital obtained on more favorable terms.[20] Again, whether the market would merely be properly readjusting its expectations in the light of new information or whether it would be irrationally attaching an unwarranted risk premium to the carrier's required performance is beside the point. There is an adverse effect on the carrier's ability to borrow funds that diversification might affect favorably. However, as in the example above, the benefits might not flow entirely in one direction. There might be times when the "industry effect" works in reverse; when a carrier would be able to secure funds on *more* favorable terms than its own performance might dictate, due to superior recent performance by the industry.[21] Loss of industry identification in the eyes of the market might cost the firm this potential advantage. Furthermore, when a carrier associates itself with a conglomerate, it assumes the risk that the market at some point in the future might, for rational or irrational reasons, discriminate against "conglomerates."[22] It also assumes the risk that poor performance by one division of the firm of which it is a member might impair the creditworthiness of the carrier, even though it might be effectively isolated from the source of the difficulties.

---

[20] Any such "industry effect" is not necessarily limited to just the regulated industry of which the firm is a member. United Air Lines claims it experienced additional difficulties in securing the funds it needed to pursue its diversification program as a result of the Penn Central collapse. CAB Docket 24283, Tr. 1605. This difficulty may be traceable, however, to the fact that the Penn Central collapse cast suspicion on diversification activities of all regulated transportation enterprises, and United Air Lines had recently announced that it was forming a holding company for the purpose of engaging in diversification. Perhaps, the relevant "industry" here was "diversifying regulated transportation firms."

[21] Again, this is not necessarily limited to firms in the industry in question. Galbraith reports that the stock of the Seaboard Air Line Railroad was a favorite during the 1920s due to the mistaken belief that the railroad was in some way connected with the then glamorous airline industry. See Galbraith [8], pp. 48–49.

However, claims of financial benefits go beyond the assertion that diversification might permit certain capital market imperfections to be overcome and attach significance to the act of diversification itself. Put most simply, the argument is that the merging of two imperfectly correlated income streams produces a less variable income stream. Thus, the joining of two companies possessing such imperfectly correlated income streams produces a reduction in risk as perceived by investors, even if the merger produces no beneficial operational effects for the combined enterprise. This reduction in risk is translated into a higher risk-adjusted rate of return which in turn should result in a lower cost of equity to the merged firm.[23]

While diversification may reduce the cost of capital in the manner just noted, it does not follow that any such benefit will necessarily flow through to the regulated portion of the enterprise. The management of the parent corporation has a duty to its stockholders to employ any available funds where they will produce the best return. Unless prospective investments in the regulated enterprise are as promising as the alternatives available to the management, additional funds will not (and should not) flow into the regulated enterprise even if they are available.

Jervis Langdon, Jr., formerly Chairman of the Board and President of the Rock Island Railroad and now a Trustee of the Penn Central, recognized this likelihood in an interview published in *Railway Age* in 1969:

> I realize that lack of capital for improvements running into the billions has a direct bearing on the railroad service problem, and that conglomerates, with their promise of greater earnings and financial strength, should be able to help in this connection. But would they really help when the choice is between a railroad investment which produces a small return and another which promises a much greater return? I wonder to what extent conglomerates, with earnings from diverse sources, can afford to pump into their railroad subsidiaries the kind of money that is required, particularly when there are so many attractive investment opportunities outside the railroad business.[24]

Besides lowering the cost of capital, diversification is said to offer the opportunity to reduce operating costs by permitting a more effective use of management talent and a coordination of related activities among subsidiaries. Examples of the latter might be the joint use of a computer facility by an airline and a hotel or the joint purchasing of coffee by the two.[25] However, one factor mitigating against the realization of significant cost savings from this source is the direct trade-off that may exist between the "operational synergy" it reflects and the "financial synergy" that produces a reduced cost of capital to the diversified enterprise mentioned above. The potential for the former is increased if the firm's various lines of business have similar operating characteristics. The latter is increased as the profit streams from the various lines of business are less directly related. While these two are not entirely mutually exclusive, the latter may require that the lines of business display somewhat dissimilar operational characteristics.

---

[22] This may already have occurred. See, for example, Lewis Berman, "What We Learned From the Great Merger Frenzy," *Fortune*, April 1973, p. 70.

[23] Weston, Smith, and Shrieves [25], p. 362, have reported results indicating that conglomerate firms have, on average, been successful in improving risk/return characteristics of the "portfolio" of firms they managed. However, they stress the importance of high quality management with competence in the "new managerial skills (planning and control, research management, information systems) as well as capabilities in the new advanced technologies."

[24] See [4].

The argument that diversification permits a more effective use of scarce management talent deserves close examination. Looked at conventionally, this seems to imply that the regulated transportation firms are subject to increasing returns to scale in management and that diversification is one means of achieving minimum optimal size. If so, there is little evidence to support such an assertion. Studies of regulated transportation firms have either found evidence of constant returns to scale or have failed to pinpoint management as the source of any economies that might have been discovered.[26] Furthermore, the possibility that the achievement of significant financial synergy requires the combining of lines of business whose operational characteristics are dissimilar also mitigates against the consolidation of very many management functions. While it has been argued that management skills are not industry specific and are easily transferred across dissimilar product lines, one survey characterized this belief as "debatable" and took note of a lack of hard evidence on either side of the question.[27]

Diversification by regulated transportation enterprises may promise significant managerial benefits of a different sort, however. Williamson has hypothesized that firms organized along functional lines (manufacturing, marketing, finance, etc.)—termed by him unitary-form, or "*U*-form," corporations—are inherently less efficient than firms organized along multidivisional lines—"*M*-form" corporations.[28] As the former firms grow in size, their managers encounter increased difficulty in enforcing cost consciousness throughout the firm because of the problem of securing accurate information about subgroup performance. As a result, elements of the "*U*-form" corporation are likely to engage in the pursuit of goals other than profit maximization—with attendant adverse effects on internal efficiency. Corporations organized along multidivisional lines avoid these problems by creating mini-firms within the firm, each of which is small enough to be managed efficiently and each of which is judged primarily upon its profit performance. In an "*M*-form" corporation, top management is freed from devoting its primary attention to day-to-day operations and can concentrate its attention instead on long-range strategic questions.

---

[25] United Air Lines reports savings from both sources as a result of its merger with Western International Hotels. See testimony of A. M. deVoursney, CAB Docket 24283, Exhibit U-T-5, pp. 11–12.

It might be argued that certain of these benefits could be achieved without an actual merger. For example, a firm might announce that it has excess computer time and offer it to the highest bidder rather than just to a sister subsidiary. While certainly possible, this suggestion ignores the fact that once two businesses are jointly managed, the transactions cost of making such an arrangement may decline substantially. See McKean [15], pp. 124–125.

[26] On trucking see Patton [21] and sources cited therein, note 12. On railroads see Moore [19], p. 67, and Keeler [14]. On airlines see Eads [6] and Testimony of James C. Miller, III, in the American-Western Merger Case, Docket 22916.

[27] DOT-DE-1, p. 20.

[28] See Williamson [26] and Chandler [3].

Williamson identifies the regulated common carriers as one of a small group of firms in the economy that have been conspicuously slow to adopt the multidivisional form of organization.[29] He attributes this in part to the fact that the nature of their product "may not lend itself to meaningful divisionalization along *M*-form lines."[30] However, Williamson also specifically identifies the reluctance of regulatory authorities to permit carriers to diversify into nonregulated areas "so as to preserve 'clean' regulatory relations" as another factor that may have inhibited the adoption of this organizational form by the regulated common carriers.

To sum up, diversification by regulated transportation enterprises indeed produces the potential for benefits that may accrue to the public at large, but is by no means clear that such benefits necessarily will be realized in practice. Diversification may result in a reduction in the cost of capital to the parent firm, but any resulting improvement in the availability of capital will accrue to the regulated portion of the business only if the prospective rate of return it can offer is attractive relative to the return that the diversified firm could earn by investing in other lines of business. Thus, diversification and any lower cost of capital it might bring will not substitute for regulatory policies that will produce a competitive rate of return for regulated enterprises. Regulatory agencies must constantly keep this in mind and also must guard against the danger that ". . . [they] (and the public) may be moved to irresponsible decisions on the premise that if the company fails to derive adequate revenues from the regulated services, it can always fall back on its unregulated revenue."[31]

Second, while cost savings may result from diversification, there is no evidence to suggest that they are likely to be great. This is particularly true if the various lines of business of the parent firm are functionally dissimilar.

---

[29] See Williamson [26], pp. 162–163.

[30] In Williamson [26], p. 163. The recent reorganization by UAL, Inc., of its United Air Lines subsidiary into geographic and functional profit centers is an attempt to see whether such a decomposition can be made. UAL, Inc., 1971 Annual Report, p. 9. This reorganization was initiated by UAL's new President, Edward E. Carlson, a man with no previous airline experience, who was formerly head of UAL's Western International Hotels subsidiary. For an indication that this decentralization is producing precisely the results that Williamson's theory would predict, see "How United Airlines Pulled out of Its Dive," *Business Week*, June 29, 1974, pp. 66–70, particularly p. 68.

[31] "Diversification by Utility Companies," Report of the Ad Hoc Committee on Nonutility Investments, published by the National Association of Regulatory Utility Commissioners, Washington, D. C., 1972, pp. 20–21.

Finally, improvements in the utilization of management talent are not likely to produce major savings, since there is little evidence to suggest that the regulated transportation firms currently are so small as to be unable to realize fully all economies of scale in this area. However, the ability to diversify may allow regulated transportation firms to adopt the form of business structure favored by the vast majority of major industrial enterprises.[32] Adoption of this structure should provide the opportunity for significant improvements in internal operating efficiency and for increased rationality in the allocation of capital.

4. Potential threats to the public interest from diversification

■ Just as successful diversification can benefit the stockholders of a regulated enterprise, an unsuccessful diversification program can injure their interests. The question is whether, to protect the public interest, regulatory agencies should have the power to supervise diversification closely or even ban it, regardless of any potential benefits.

We have noted above that the principal concern that led to the enactment of the "commodities clause" was the fear that entry by railroads into other lines of business might have permitted the extension of whatever monopoly power they then possessed into otherwise competitive industries. The fears that have been expressed concerning the recent wave of railroad diversification reflect an almost 180-degree shift in opinion. According to the staff of the ICC:

> The present threat to the public interest is not the creation of railroad monopolies or restriction of intra-railroad competition, but the dilution of the interest of managements of the conglomerate in the provision of efficient transportation services and the potential dilution of the capital structure of the railroads. More specifically, the conglomerate holding company provides a convenient means for the transfer to other industries of assets now devoted to transportation.[33]

Put slightly differently, the ICC staff believes that if railroads are permitted to diversify, there is a strong possibility that they will use this opportunity to move assets out of railroading and into other lines of business.

During the course of its investigations of railroad diversification activities, the ICC staff did uncover a number of cases of asset transfer that had been carried out under circumstances it considered highly irregular. Railroads were found to have paid dividends to parent holding companies in excess of levels that appeared prudent. In some cases these dividends consisted of properties whose true values substantially exceeded the nominal values at which the transactions were recorded. Holding companies were found to have charged railroads fees in excess of the value of services provided and to have borrowed money from railroads at extremely favorable rates of interest. A study carried out by the staff of the Senate Commerce Committee concluded:

---

[32] According to Scott [23], "By 1967 over 90 percent of the Fortune "500" were diversified to some extent, and over 80 percent were managed by the multidivisional firm," pp. 138–139. Since 1967 both percentages have increased. See Exhibit III, p. 139.

[33] See ICC [11], pp. 1–2.

> [w]here a railroad holding company or linked firm relationship exists, the opportunities for abuse are limited only by the ingenuity of their financial managers, lawyers, and accountants.[34]

To correct such "abuses" the ICC staff recommended that the Commission reconsider its previous position that it lacked jurisdiction over railroad holding companies. The staff recommended that the Commission should seek legislation granting it jurisdiction if it concluded that such jurisdiction was lacking. The staff also recommended that the Commission develop regulations to control the transfer of assets out of railroading into other lines of business. One specific recommendation was that the Commission require its prior approval of *all* transactions between a railroad and its affiliates as being consistent with the public interest.[35]

It is indisputable that diversification by railroads has resulted in a transfer of assets out of railroading into other lines of business, statements of railroad management to the contrary notwithstanding. But it is also clear that the mere fact that other activities engaged in by regulated firms utilize assets that once may have been employed by the regulated carrier does not *by itself* constitute a sufficient reason for regulating such asset transfers. As the Department of Transportation has observed:

> There is nothing intrinsically evil, or illegal, in these types of transactions. The movement of assets within a diversified corporation is one of the primary tools available to management to maximize the return on shareholders' investment.[36]

What may create a public interest in the regulation of such asset transfers is the possibility of a conflict between the managerial goal of maximizing the overall return to the stockholders and the fulfillment of the public service obligations of the regulated carrier—at least as these obligations are viewed by the regulatory agency.

It is worth considering how such a conflict might arise. Posner has suggested that one major goal of regulatory agencies has been to expand service availability beyond the point that the market will support, and to finance such an expansion by cross subsidization.[37] Successful cross subsidization requires that profits in excess of a normal rate of return be earned in some portion of the business. This has generally been accomplished by adopting some means of discriminatory, or "value-of-service," pricing. Under such a scheme, customers whose demand is relatively inelastic are called upon to bear a larger share of the common costs and contribution to profit than are the customers whose demand is relatively elastic.[38] While this may prove to be a successful means of financing service expansion in the short run, it is not likely to prove viable in the long run. Technological change is likely to be stimulated precisely in those areas where demand is most inelastic, for the potential payoff to innovation is greatest there. Potential competitors are also likely to look to those customers who are being most discriminated against as targets for the development of new services. Consequently, the potential for "cream skimming"—the entry of new competitors into the relatively profitable lines of the regulated firm's business—is almost certain to appear eventually. This places the regulatory agency in the position of being the defender of monopoly in the face of clearly demonstrable opportunities for competition. In some cases the agency may decide to permit competition to develop. In other cases the agency may attempt to limit the ability of the new competition to erode the excess profit base by bringing it under regulation.[39] But even if the regulatory agency adopts this latter posture, as the ICC did with respect to common carrier trucking, it is not likely to be totally successful in thwarting competition. As long as prices are held above the level that would produce at least normal profits for the most efficient potential competitor, the pressures for competition will remain.

---

[34] See [1], p. 77. The report continued: "The cumulative effect of these practices can be to siphon cash from the railroad and weaken its capacity to function as a common carrier."

[35] See [11], pp. 954–956, 1044–1045. The ICC has requested Congress to extend its jurisdiction to cover railroad holding companies. See *Railroad Industry Overview—1971*, pp. 326–330.

[36] Testimony of Russell F. Murphy, DOT-T-1, p. 37. On p. 25 of his testimony, Mr. Murphy states that one positive effect of diversification is the possibility it creates for "thawing frozen assets." He cites as one example "the transfer by railroads of unimproved property acquired in the land-grant era through their parent holding companies to a real estate subsidiary."

[37] See Posner [22].

Once monopoly revenues begin to erode, the regulatory agency is faced with a series of alternatives, all of which may have unfortunate consequences from its point of view. First, the agency can allow the regulated firm to drop enough unremunerative services so that its revenues and costs can be brought into line. This is certain to provoke a public outcry by those formerly receiving the benefits of the implicit subsidy and is likely to generate Congressional wrath. Second, the agency can request an explicit subsidy from Congress to offset the loss of monopoly revenues. This may also be unpalatable, and it is sure to attract opposition from the Office of Management and Budget. Finally, the agency can deny the firms it regulates permission to abandon unprofitable services. Doing this guarantees that these firms will be unable to earn the rate of return they require to attract capital, and hence causes their eventual contraction or extinction.

The ability of regulated firms to diversify may deny this third option to the regulatory agency and set up an apparent conflict between the "public interest" and the interest of the firm. Through a judicious use of "creative accounting," the regulated firm may be able to withdraw assets from the regulated sector of the business while maintaining the appearance of continuing to offer the unremunerative services. The low level of financial expertise at the regulatory agencies, the weakness of agency enforcement efforts, and the disinclination to move aggressively against a regulated carrier unless abuse of the agency's will becomes blatant contribute to the likelihood that this will occur.

---

[38] Certain economists do not condemn all uses of value-of-service pricing. If an industry is subject to increasing returns to scale, some form of value-of-service pricing may be the least burdensome means by which a public utility can finance its operations. See Baumol and Bradford [2].

[39] This reflects what Kahn has referred to as "the tendency of regulation to spread," [13], p. 28.

In the case of the railroads, the redeployment of assets out of the regulated sector is clearly in the public interest.[40] Yet, it seems likely that if a redeployment of assets can take place only in this "backdoor" manner, it may not proceed optimally. Services may be curtailed not where they are least needed by the public, but where their curtailment is least likely to provoke an embarrassing outcry in the short run. Assets that may be sold to finance diversification may not be those yielding the lowest return when employed in the regulated portion of the enterprise, but rather those whose sale can be accomplished with the lowest risk of attracting public notice. Finally, if the service burdens imposed by the regulatory agency are severe enough to have forced the carrier close to bankruptcy, management may engage in a short-sighted program of selling assets to raise the cash required to maintain operations.

Clearly, such nonoptimal redeployment of assets might impair the firm's ability to carry on the regulated portion of its business. In an extreme case, it might result in bankruptcy of the common carrier subsidiary (though not necessarily of the parent holding company). It might be argued that these possible consequences provide a sufficient justification for close control of regulated carrier diversification activities. However, such an argument is not convincing. In the case of the railroads at least, it is precisely the reluctance of the regulatory agency to face up to the consequences of the failure of its regulatory strategy that has created the primary *incentive* for diversification. If regulation is the cause, more regulation hardly seems the proper cure, for it is highly unlikely that an ICC with authority to regulate carrier disinvestment would abandon its previous policy of resisting needed adjustments in the scale of railroad operations. Instead, the Commission probably would use its authority to bar disinvestment as a means of escape from its regulatory authority. Alternatively, the Commission might attempt to direct the course of diversification to conform with its view of the public interest.

The adoption of either of these two courses would involve the Commission deeply in the internal management affairs of the railroads. As the ICC staff noted, in order to determine "when disinvestment may be justified or when it turns into unjustified dissipation of assets," it might be necessary for the Commission to substitute its judgment for that of the management.[41]

---

[40] See *The Northeast Railroad Problem*, A Report to the Congress submitted by the Secretary of Transportation in response to S.J. Res. 59-2, U. S. Department of Transportation, March 26, 1973. Moore [19], pp. 79–81, estimates the deadweight loss to the economy resulting merely from the inefficient use of the railroads as between $1.7 billion and $2.4 billion. Much of this is attributed to excess capacity generated (and maintained) by regulation. See also Harbeson [8], Friedlaender [7], chapter 4, and Meyer *et al.* [16], Chapter 6. Even the staff of the ICC appears to agree that some movement of assets out of railroading into other lines of business is desirable. See [11], p. 3. But the ICC's reaction to the Penn Central bankruptcy and to the Administration's plan to rehabilitate the railroad by paring down its size considerably might cause one to question the depth of the Commission's belief in this proposition. See John E. Cooney, "The Northeast's Big Railroad Muddle," *The Wall Street Journal*, Tuesday, June 12, 1973, p. 26.

Given the Commission's past record in interpreting the public interest, it is difficult to be sanguine about the results. Furthermore, such an intrusion into what are normally considered the prerogatives of management would seriously impair the ability of the railroads to attract and hold management talent of high quality.

5. Evidence on the effects of diversification

■ If we disregard the rhetoric that appears to consider it either unethical or immoral for assets to be transferred out of railroading without "full compensation" paid the carriers,[42] or, indeed, for management even to consider the possibility of moving assets out of railroading, and instead attempt to determine what effect railroad diversification has had on the carriers, the picture is not so dark as some might have us believe. While there are few cases that can be found where the railroad has clearly benefitted from diversification,[43] there is no solid evidence to suggest that the railroads have been seriously harmed by their association with holding companies, despite some asset removal.

As noted earlier, there have been two cases in which railroads formed holding companies for the purpose of engaging in diversification and then proceeded to diversify entirely out of railroading. The ICC staff makes much of the "financial abuses" committed by the Bangor Punta management, and claims that it owes the railroad

---

[41] See [11], pp. 40–41.

[42] It is not entirely clear what the advocates of "full compensation" have in mind. Some assets may have one value when employed in railroading and another quite different value when employed elsewhere. For example, loss carryforwards in excess of a railroad's ability to make use of them may be of little or no value to the railroad standing alone, but can be utilized by a profitable business acquired by the railroad and given a value during merger negotiations between the two. But to what extent "ought" a carrier to benefit from such assets? Management policies vary widely. In the Air Carrier Reorganization Investigation, Braniff Airways stated that once it formed its holding company, its policy would be to grant Braniff full and immediate credit for any tax advantages created by the carrier's association with the group. See Proposed Findings of Fact and Conclusions of Braniff Airways, Incorporated, and Braniff International Corporation, Air Carrier Reorganization Investigation, Docket 24283, pp. 37–38. UAL, Inc., stated that its policy was to grant its airline subsidiary full credit at the time the carrier could make use of any tax benefits up until the time that the credit would expire. See testimony of Irvin E. Williamson, Vice President, Finance, UAL, Inc., Exhibit U-T-6, pp. 15–16. However, when a bankrupt subsidiary of the Norfolk and Western Railway asked the N&W to give back some of the tax benefits the N&W received because of the bankrupt carrier's association with it, the N&W stated that its counsel had advised that: "the Erie trustees' arguments to have income tax benefits returned to the Erie and to have the Erie severed from the accounts of the profitable parent are without legal merit." See "N&W Says Subsidiary, Erie Lackawanna, Asks Return of Tax Benefits," *The Wall Street Journal*, Thursday, May 17, 1973, p. 28.

[43] The ICC staff concluded that the Bangor and Aroostook Railroad received no benefits from its association with Bangor Punta. See ICC [10], pp. 3–8. One possible benefit reported by the DOT study of five railroad holding companies was the CN&W's ability to borrow substantial sums at prime interest rates from Northwest Industries during 1969, the year when it faced its most serious financial problems. See DOT-DE-3, p. 41.

$64 million.[44] However, nowhere in its report does the staff suggest that the transferal of assets actually impaired the ability of the Bangor and Aroostook to operate efficiently. A DOT study of the effects of the diversification by Northwest Industries on the Chicago and Northwestern Railway concludes that although the railroad was undermaintained and in at best "marginal" financial condition when disposed of by Northwestern Industries, there was no evidence to suggest that its condition would have been any better if the diversification had never taken place.[45] And after examining the diversification activity of five railroads, the DOT concluded:

> In none of the cases studied did publicly available information indicate that the formation of a holding company and its subsequent control of the railroad have any negative effects on the railroad's performance or on reinvestment in the railroad.[46]

6. Conclusions

■ Norton Simon, the noted industrialist and conglomerateur, recently organized a conference in New York City to offer his views on proposals to solve the current "rail crisis." One of Simon's criticisms of these proposals was that they "would let the railroad conglomerates like the Union Pacific concentrate on drilling for oil in the North Sea when they should concentrate on running the railroad."[47] These views echoed the warnings of the ICC staff that one of the principal threats to the public interest resulting from diversification by railroads is "the dilution of the interest of management of the conglomerates in the provision of efficient transportation services."[48] Indeed, the view is widespread that it is somehow appropriate for firms like Norton Simon, Inc., to diversify widely[49] and to enter and leave lines of business based solely upon judgments as to profitability,[50] but it is inconsistent with the public interest for railroads to do likewise. This belief goes to the heart of the question of whether railroad diversification is in the public interest even though the potential for "asset dissipation" is an inevitable by-product.

---

[44] See ICC [10], Appendix VIII.

[45] See DOT-DE-3, pp. 41–44.

[46] See DOT-T-1, p. 3. DOT did stress, however, that it had only used publicly available information. See DOT-T-1, p. 4.

[47] See "Norton Simon's Do-It-Yourself Railroad Plan," *Business Week*, November 17, 1973, p. 74.

[48] See ICC [11], pp. 1–2.

[49] In 1972 the sales of Norton Simon Industries (which Mr. Simon left five years ago to devote his time to educational and cultural interests) were distributed by major product line as follows:

Food and Food Services	43.7%
Communications and Graphic Systems	21.3%
Soft Drinks	10.2%
Distilled Spirits	11.8%
Packaging Systems	13.0%

Source: [17], p. 2811.

[50] For a listing of the industries entered and left by Norton Simon, Inc., since the firm was formed in early 1968, see [17], p. 2810. For an account of Mr. Simon's own entry into and exit from the steel industry, see Dan Cordtz, "Antidisestablishmentarianism at Wheeling Steel," *Fortune*, July 1967, p. 105.

One basis for the "management inattention" argument is the fear that a railroad management pursuing a diversification program would inevitably fail to devote "sufficient" attention to the affairs of the railroad. However, we know that there is no reason why the efficient operation of a railroad should demand that top management concern itself only with the details of day-to-day railroad operations. We noted above that the multidivisional form of corporate organization is specifically designed to free top management from such tasks while still maintaining maximum incentive for efficiency in operations. However, we also noted that regulated common carriers have been among the slowest firms to adopt this important organizational innovation and that this may have been due in some measure to the desire of the regulatory agencies to preserve "clean" regulatory relations by denying the firms they regulate permission to diversify. Consequently, diversification and the organizational changes that are likely to accompany it should offer the promise of long-run improvements in railroad operating efficiency through a reduction in organizational slack. This would be of benefit to the railroads and would almost certainly be applauded by the ICC.

However, the reorganization of railroad companies along multidivisional lines as a consequence of diversification would produce other results that the ICC would be less likely to applaud. This may be thought of as the other side of the "management inattention" argument. The same factors that tend to make corporations organized along multidivisional lines more efficient in their day-to-day operations also tend to make such firms more rational in the allocation of capital to their various component parts than firms organized on a unitary basis. Just as inefficient, slack-creating operating procedures are less likely to be tolerated by an *M*-form corporation, so the tendency of any subunit to invest in activities that fail to promise a risk-adjusted rate of return comparable to that which the firm could earn by investing in its other lines of business is likely to be strongly curbed. Indeed, according to Williamson, "a profit-oriented general office in an *M*-form enterprise might be expected easily to secure superior performance to that which the capital market can enforce."[51]

Ordinarily such an improvement in capital allocative efficiency would be considered a major benefit both to the stockholders of the enterprise and to the public. After all, capital is a scarce good, and economy in its use ought to be encouraged. Unfortunately, a higher degree of rationality in allocating capital among subunits is one of the things that the ICC apparently fears may result from railroad diversification. As one ICC member stated in an interview:

> . . . the very reasons which have stimulated the trend toward transportation conglomerates and the asserted benefits suggest some potential dangers which cannot be ignored by the Commission. A principal reason for diversification has been the low rate of return on railroad investment as compared to the greater profits available in nontransportation enterprises and the fact that income from nontransportation enterprises would provide additional financing to maintain and upgrade rail plants. There are many who ask, however: would there be more incentive or even pressure on top management to

---

[51] See Williamson [26], p. 140.

reinvest such income in businesses which will produce a greater rate of return than the railroads? Without strong public service motivation, would this not be the most predictable course to follow?[52]

Apparently, in this Commissioner's view, and perhaps in the view of Norton Simon, a railroad management that demanded a compensatory rate of return before it would invest in any given line of business, that would reduce the level of assets it is willing to commit to any line of business whose profitability is declining, and that would even be willing to sell off a line of business to someone else or liquidate it if its future prospects were not sufficiently promising, would be a railroad management lacking "strong public service motivation."

It is precisely because both railroad managements and regulators have believed that railroading is somehow different from other businesses that the railroads find themselves in their present plight. If diversification will contribute to an ending of this belief, then it should be strongly encouraged. Our recommendation is not changed by the realization that in the short run, diversification will probably result in assets' flowing out of railroading, perhaps in a nonoptimal fashion, as railroad managements attempt to place their resources where they can earn a competitive rate of return for their stockholders. Any costs associated with the nonoptimal nature of this adjustment must be considered as yet another of the costs of railroad regulation that already appear to be so high.[53] Furthermore, granting the ICC extensive powers to regulate this redeployment of railroad assets will not assure that it will proceed in a manner more consistent with the public interest, but instead is likely either to halt or to hamper this much needed adjustment. At worse, expanding the Commission's powers may severely aggravate the problems that the railroads currently face in attracting and holding high quality management, a problem that Mr. Simon also noted in his recent remarks.[54]

One reason for examining railroad diversification was to shed some light on problems that might emerge elsewhere. In other regulated industries, such as telecommunications, the postal service, and the airlines, highly discriminatory rate structures have been erected with the approval (and often at the insistence) of the responsible regulatory authorities, and the additional profits generated have been utilized to support an expansion of service beyond the level that the market would justify. These discriminatory rate structures have created opportunities for "cream skimming" by competitors. Certain of the regulatory agencies, far from seeking to bar the emergence of competition, have tended to encourage it.[55]

Most economists (including this one) have applauded such moves in the direction of allowing more competition. However, what must be realized by regulators and economists alike is that a decision to permit competition must be accompanied by a decision about how (and at what level) to support "public interest" services once the revenues that previously financed them dry up with the introduction of competition. Either direct subsidies must be voted by Congress or the firms newly experiencing competition must be allowed to adjust the scale of their "public interest" services so that the costs of these services can be covered by the reduced revenues. Otherwise pressures will be created that may force the affected firms to consider methods of redeploying their assets in ways that may be more costly to society. There is nothing unique about the railroads in this regard.

---

[52] See [4], p. 36.

[53] See Moore [19], pp. 79–81.

[54] See note 47.

[55] For a review of the history of the FCC's policy shift on competition, see Kahn [13], pp. 127–145.

References

1. *The American Railroads: Posture, Problems, and Prospects.* Staff Analysis for the U. S. Senate Committee on Commerce, Prepared at the Direction of Honorable Warren G. Magnuson, Chairman, Committee on Commerce, United States Senate, August 28, 1972.
2. Baumol, W. and Bradford, D. "Optimal Departures from Marginal Cost Pricing." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp 265–283.
3. Chandler, A. D., Jr. *Strategy and Structure.* New York: Doubleday, 1966.
4. "Conglomerates: How Will They Affect Rail Service?" *Railway Age*, April 21, 1969, p. 36.
5. Daughen, J. R. and Binzen, P. *The Wreck of the Penn Central* New York: Mentor, 1973.
6. Eads. G. "A Cost Function for the Local Service Airlines." Unpublished Ph.D. Thesis, Yale University, 1968.
7. Friedlaender, A. *The Dilemma of Freight Transport Regulation.* Washington, D. C.: The Brookings Institution, 1969.
8. Galbraith, J. K. *The Great Crash.* Boston: Houghton Mifflin Co., 1954.
9. Harbeson, R. W. "Toward Better Resource Allocation in Transport." *Journal of Law and Economics,* Vol. 12, No. 2 (October 1969), pp. 321–338.
10. Interstate Commerce Commission, Bureau of Accounts. "Review of Diversified Holding Company Relationships and Transactions of Bangor Punta Corporation." February 1971.
11. Interstate Commerce Commission, Bureau of Economics. "Conglomerate Merger Activity of Class I Railroads and Related Regulatory Problems." March 1969.
12. Interstate Commerce Commission. "Conglomerate Merger Activity Involving Motor Carriers." October 1970.
13. Kahn, A. E. *The Economics of Regulation*, Vol. II. New York: John Wiley, 1971.
14. Keeler, T. E. "Railroad Costs, Returns to Scale, and Excess Capacity: A Neoclassical Analysis." Working Paper No. 35, Department of Economics, University of California, Berkeley, April 1973.
15. McKean, R. "Discussion of O. E. Williamson's 'The Vertical Integration of Production: Market Failure Considerations.'" *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 124–125.
16. Meyer, J. R. et al. *The Economics of Competition in the Transportation Industries.* Cambridge: Harvard Univ. Press, 1959.
17. *Moody's Industrial Manual*, 1973 edition.
18. *Moody's Transportation Manual*, 1972 edition.
19. Moore, T. G. *Freight Transportation Regulation.* Washington, D. C.: The American Enterprise Institute for Public Policy Research, 1972.
20. National Association of Regulatory Utility Commissioners. *1972 Report of the Ad Hoc Committee on Non-Utility Investments—Diversification by Utility Companies.* Washington, D. C., 1972.
21. Patton, E. M. "Implications of Motor Carrier Growth and Size." *Transportation Journal* (Fall 1970), pp. 34–52.

22. POSNER, R. A. *"Taxation By Regulation." The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.
23. SCOTT, B. R. "The Industrial State: Old Myths and New Realities," *Harvard Business Review* (March-April 1973), pp. 138–139.
24. UNITED STATES CONGRESS. *The Fenn Central Failure and the Role of Financial Institutions.* Staff Report of the Committee on Banking and Currency, House of Representatives, 92nd Congress, 1st Session, January 23, 1972.
25. WESTON, J. F., SMITH, V. K., AND SHRIEVES, R. E. "Conglomerate Performance Using the Capital Asset Pricing Model." *Review of Economics and Statistics*, Vol. 54, No. 4 (November 1972), pp. 357–363.
26. WILLIAMSON, O. E. *Corporate Control and Business Behavior.* Englewood Cliffs, N. J.: Prentice-Hall, 1970.

The losses on savings deposits from interest rate regulation

David H. Pyle

Associate Professor
School of Business Administration
University of California, Berkeley

Ceilings on the deposit rates payable by savings institutions under the Interest Rate Adjustment Act of 1966 resulted in interest income losses for savers. A recursive, rate-adjustment model of deposit rate formation is used to predict the deposit rates which would have been paid in the absence of the Act. The resulting estimate of lost interest income during the 1968 through 1970 period is over five billion dollars.

1. Introduction

■ During the mid-1960s, the Federal Reserve used deposit rate regulation under Regulation Q to attempt to reduce the growth of bank credit. By the *Interest Rate Adjustment Act of 1966*, the Congress extended maximum rate regulation to the other major savings institutions and established a positive spread between the maximum rates payable by these other institutions and the maximum rates payable on comparable savings deposits at commercial banks. This extension of maximum rate regulation resulted from a concern regarding the solvency of the nonbank savings institutions in a period of rapidly rising short-term interest rates and a concern over the supply of funds to the mortgage market.

As a result of the regulation of the maximum interest rates payable at savings institutions, savers who held interest-bearing accounts in commercial banks, mutual savings banks, and savings and loan associations received less interest income than they otherwise would have received. The objective of the research reported here was to estimate the size of the opportunity losses that resulted from this regulation.[1]

In the analysis that follows, the effects of large certificates of deposit (CDs) have not been included in the loss estimates. Therefore, it seems safe to suggest that the estimated losses were incurred

---

Funds for this research were provided by a Dean Witter Foundation Grant. Computer time was provided by the UC Computer Center. This assistance is gratefully acknowledged. This paper has been improved by comments on an earlier draft by members of the Stanford-UC Finance Colloquium, by other colleagues at Berkeley, including in particular Charles Kelso and Barr Rosenberg, and by comments from Fischer Black and Paul MacAvoy. Research assistance was provided by Jay Morrisson.

[1] See Pye and Young [4] for a discussion of the macroeconomic effects of ceilings. The authors conclude that ceilings are not neutral, including their effect on interest rates. I have not attempted to include this effect in the calculation of losses to savers. For a discussion of the policy implications of deposit rate ceilings, see Kane [1]. The research reported here could be thought of as a quantification of the discrimination discussed by Kane.

by savers who did not find suitable alternative ways of saving. Alternatives offering substantially higher interest rates than those available from savings institutions were available to some savers; the shift of corporate short-term investments from CDs to commercial paper is an illustration of the behavior of these savers. (See Table 1.)

TABLE 1

ANNUAL RATE OF GROWTH (IN PERCENT) IN THE STOCK OF CDs, COMMERCIAL PAPER, AND SAVINGS DEPOSITS, 1967-1970

	CDs	COMM. PAPER	SAVINGS DEPOSITS
1967	32.1	24.3	11.6
1968	14.5	23.9	8.6
1969	−53.0	54.6	2.5
1970	132.4	0	8.9

In contrast, consider the effects of maximum rates on other types of savings deposits during the same period.[2] The compound rate of growth of savings deposits during 1960–1965 was 11.3 percent. From Table 1, we see that 1967 was a typical year in this respect. Effective maximum rates in 1968, 1969, and 1970 resulted in a compound deposit growth rate over this three-year period of 6.6 percent. Even in 1969, however, when there was a massive outflow of CDs, the savings deposit growth rate remained positive. Why did savings deposits not leave savings institutions at a rate similar to that experienced by commercial banks on CDs? Presumably because many savers did not know of or could not avail themselves of suitable alternatives offering higher rates. Transaction costs, minimum transaction quantities, information costs, and inconvenience prevented savers from buying short-term market securities.[3] Most importantly in terms of the losses reported here, interest regulation under the *Interest Rate Adjustment Act of 1969* precluded interest-rate competition among the savings institutions that had previously provided low transaction cost, low minimum balance, convenient alternatives for these savers.[4]

---

[2] The data on average savings deposit rates are from the *1973 Savings and Loan Fact Book* [5], p. 17, except for the commercial bank rates after 1964. Due to significant interest payments on negotiable CDs after 1964, the data in the *Fact Book* are unsatisfactory for the purposes of this research. In 1965 and 1966, the rate on savings deposits at commercial banks was calculated by deducting interest payments on business deposits from total interest and dividing by non-CD average deposits. From 1967 on, a weighted average of the most common interest rates paid on savings deposits and time deposits of less than $100,000 taken from Federal Reserve surveys was used. As implied by these rate definitions, savings deposits in this paper are taken to be savings capital at savings and loan associations, time and savings deposits other than large CDs at commercial banks, and deposits at mutual savings banks.

[3] On February 25, 1970, the minimum noncompetitive bid for Treasury bills was increased from a bid for $1,000 in bills to a bid for $10,000 in bills. See Mullineaux [3] for evidence regarding the resulting reduction in disintermediation.

[4] The interest rate ceilings under the Act became effective on September 26, 1966.

2. The losses on savings deposits

■ To estimate the size of the opportunity losses to savers as a result of the *Interest Rate Adjustment Act*, an estimate of the interest rates that would have been paid on savings deposits in the absence of this law is needed. Note that the emphasis here is not on the effects of maximum deposit rates *per se*, since commercial banks were subject to ceiling rates under *Regulation Q* prior to the passage of the *Interest Rate Adjustment Act*. Rather, the emphasis is on the particular effects of the joint regulation of maximum rates at all major savings institutions.

The evidence from the growth rate of savings deposits and from simple regressions of the average interest rates paid on deposits on market interest rates suggests that the rates paid by savings institutions in 1967 were not significantly affected by the 1966 legislation. Consequently, this observation was used in obtaining the prediction equations.[5]

To obtain a set of prediction equations for the annual average rates paid by the three major institutions which offer savings deposits, a simultaneous, rate-adjustment model was formulated. The logic of the adjustment model is that the current average rate paid by a savings institution is its most recent average rate adjusted by the difference between the most recent rate and the current rate on a primary security (e.g., U. S. bonds) and by a difference between the most recent rate and the current rate on a competitive secondary security. A constant term was included in the adjustment model and may be interpreted as the normal or steady-state spread between the rate being estimated and the rate on the primary security plus the normal or steady-state spread between the rate being estimated and the rate on the competitive secondary security. A dummy variable (0 before 1961 and 1 thereafter) was included in the equation for the rate paid at commercial banks. It was intended to capture the effect of a change in the competitive attitudes of commercial banks with respect to savings deposits. The introduction of negotiable CDs in 1961 is an objective indicator of this shift in bank behavior.

The following model was estimated for the period 1952 through 1967 using two-stage least squares:

$$\begin{aligned} i_{SL}(t) = i_{SL}(t-1) &+ a_1[i_G(t) - i_{SL}(t-1) + k_{SL}] \\ &+ a_2[i_{CB}(t) - i_{SL}(t-1) + l_{SL}] \qquad (1) \end{aligned}$$

$$\begin{aligned} i_{CB}(t) = i_{CB}(t-1) &+ b_1[i_G(t) - i_{CB}(t-1) + k_{CB}] \\ &+ b_2[i_{SL}(t) - i_{CB}(t-1) + l_{CB}] + b_3 d(t) \qquad (2) \end{aligned}$$

$$\begin{aligned} i_{MS}(t) = i_{MS}(t-1) &+ c_1[i_G(t) - i_{MS}(t-1) + k_{MS}] \\ &+ c_2[i_{CB}(t) - i_{MS}(t-1) + l_{MS}], \qquad (3) \end{aligned}$$

where i_{SL}, i_{CB}, and i_{MS} are annual average rates of interest paid on savings and time deposits (other than large CDs) by savings and loan associations, commercial banks, and mutual savings banks, respectively; i_G is the annual average yield to maturity on five-year U. S. government bonds; d is a dummy variable which is 0 before 1961 and 1 thereafter; and k and l are constants in the various rate adjustment functions.

---

[5] The predictions are not significantly altered if this observation is dropped.

The results of the experiment with this model suggested that the average rate paid on the competitive secondary security (i_{CB}) was not a significant determinant of the average rate paid by savings and loan associations ($\hat{a}_2 = 0.034$ with a t-statistic of 0.169), and that the average yield on the primary security (i_G) was not a significant determinant of the average rate paid by mutual savings banks ($\hat{c}_1 = 0.0089$ with a t-statistic of 0.166)[6] Elimination of these variables reduced the rate adjustment equations to a recursive system, which was then reestimated by using ordinary least squares. The resulting prediction equations fitted on the 1952–1967 data are as follows (t-statistics in parentheses below the coefficient estimates):

$$\begin{aligned} i_{SL}(t) &= \underset{(1.72)}{0.202} + \underset{(13.94)}{0.843}\, i_{SL}(t-1) + \underset{(3.30)}{0.134}\, i_G(t) \qquad \text{(1a)} \\ R^2_{adj} &= 0.988 \qquad S.E. = 0.077 \end{aligned}$$

$$\begin{aligned} i_{CB}(t) &= \underset{(-3.13)}{-1.396} + \underset{(2.20)}{0.348}\, i_{CB}(t-1) + \underset{(2.15)}{0.131}\, i_G(t) \qquad \text{(2a)} \\ &\qquad + \underset{(3.14)}{0.700}\, i_{SL}(t) + \underset{(2.53)}{0.257}\, d(t) \\ R^2_{adj} &= 0.994 \qquad \text{S.E. } 0.084 \end{aligned}$$

$$\begin{aligned} i_{MS}(t) &= \underset{(3.79)}{0.935} + \underset{(2.22)}{0.394}\, i_{MS}(t-1) + \underset{(3.56)}{0.466}\, i_{CB}(t) \qquad \text{(3a)} \\ R^2_{adj} &= 0.992 \qquad \text{S.E.} = 0.074. \end{aligned}$$

Before reporting the predicted interest rates for 1968, 1969, and 1970, let us examine the prediction equations. First, note that each of the constant terms has a sign that is consistent with the logic of the rate adjustment model and historical interest rate spreads. In the savings and loan equation, the implied steady-state spread between the average deposit rate and the average yield on five-year bonds is about 150 basis points. The use of this result and of the constant in the commercial bank equation gives an implied steady-state spread between the average rate paid by savings and loan associations and the average rate paid by commercial banks, which is 140 basis points before 1961 and 110 basis points thereafter. In the case of mutual savings banks, the implied steady-state spread between their average rate paid and the average rate paid at commercial banks is positive as expected, but the magnitude of 200 basis points is larger than expected.[7]

Given the rate adjustment specification, the coefficient for the lagged dependent variable plus the coefficient for the competitive

---

[6] Complete results of the two-stage regressions are available from the author.

[7] By dropping the lagged dependent variable, one obtains the following prediction equation for the interest rate paid by mutual savings banks:

$$\begin{aligned} i_{MS}(t) &= \underset{(25.43)}{1.47} + \underset{(35.24)}{0.755}\, i_{CB}(t) \\ R^2_{adj} &= 0.989 \qquad \text{S.E.} = 0.084. \end{aligned}$$

The predictions for 1968, 1969, and 1970 obtained from this equation are on the average somewhat larger than those reported later for the rate adjustment equation.

secondary security rate (if included) and the coefficient for the primary security rate (if included) should sum to one for each prediction equation. In the savings and loan equation, they sum to 0.977, in the commercial bank equation to 1.179, and in the mutual savings bank equation to 0.860.

While these results appear to be reasonably consistent with the model specification, the coefficients may be biased due to serial correlation of the errors in the autoregressive equations that were estimated.[8] On the basis of sample statistics, the asymptotic bias in the coefficients appears to be small, because the estimated regression errors are very small relative to the variability of the exogenous variables over the time period.[9] For example, in the savings and loan equation, the ratio of the variance of the residuals to the variance of the bond yield is approximately 0.007. Given the small sample size, this may be cold comfort with respect to the actual biases.

Any coefficient bias that may exist does not mean there will be bias in the predictions. Perhaps the best test of the reliability of the prediction equations is to examine their performance in 1971 and 1972 after market interest rates had fallen from their 1968 through 1970 levels. Suppose that the maximum interest rates payable in 1971 and 1972 permitted savings institutions to pay interest rates consistent with the rate adjustment equations (1a), (2a), and (3a) and that the rate adjustment was based on the actual average rate paid in 1970.[10] The resulting predictions and the actual rates paid are given in Table 2. Except for the commercial bank rates (on which a discussion follows), note that the predicted rates are lower than the actual rates in three of four cases.[11] Given the estimated regression standard errors of 0.077 for $\hat{\imath}_{SL}$ and 0.074 for $\hat{\imath}_{MS}$, these predictions for the savings and loan rate and for the mutual savings

TABLE 2

PREDICTED ($\hat{\imath}$) AND ACTUAL AVERAGE INTEREST RATES (IN PERCENT) AT SAVINGS INSTITUTIONS WITH PREDICTION BASED ON ACTUAL VALUES OF INDEPENDENT VARIABLES

	$\hat{\imath}_{SL}$	i_{SL}	$\hat{\imath}_{CB}$	i_{CB}	$\hat{\imath}_{MS}$	i_{MS}
1971	5 27	5 33	5 02	4 74	5 12	5.13
1972	5 48	5 40	5.06	4 74	5.17	5 22

---

[8] The Durbin-Watson statistics (which are biased towards 2) were 2.67 for the savings and loan equation, 1.69 for the commercial bank equation, and 1.97 for the mutual savings bank equation.

[9] See, for example, Malinvaud [2], pp. 458–67.

[10] For example, the prediction for the savings and loan rate in 1971 given these assumptions is given by

$$\hat{\imath}_{SL}(1971) = 0.202 + 0.843\, i_{SL}(1970) + 0.134\, i_G(1971),$$

where $\hat{\imath}$ is a predicted rate, and i is an observed rate.

[11] It is possible that there is a downward bias in some of the coefficients due to errors in the independent variables. In particular, the five-year government rate is a proxy for the "true" primary market rate which governs the behavior of savings institutions.

bank rate seem reasonably precise. The estimated regression standard error for $\hat{\imath}_{CB}$ was 0.084.

The apparent lack of precision in the 1971 and 1972 predictions of the commercial bank rate can be explained by the fact that the rate adjustment mechanism was not feasible for commercial banks in these years. As noted earlier, federal administration of interest rate ceilings at savings institutions since the enactment of the *Interest Rate Adjustment Act of 1969* provided for a positive spread between the rates payable at the two other savings institutions and the rates payable at commercial banks. During 1967 through 1972, this spread was either 50 basis points (on all regular or savings accounts) or 25 basis points (on all other savings and time deposits with which we are concerned). The lower ceiling rates applicable to commercial banks continued to constrain their behavior in 1971 and 1972. The result is that predicted rates are greater than actual rates in these years as well as in the 1968–1970 period. In fact, the average maximum rate payable (where the average is over deposit types) was 4.85 percent in 1971 and 4.93 percent in 1972—lower than the predicted rates.[12] Legally, the predicted rates at commercial banks could not have been paid in these two years.

Given this evidence on the reliability of the prediction equations, let us turn to the predicted rates for 1968, 1969, and 1970. With the observed values for primary market rates and the predicted values for secondary market rates, the recursive system given by equations (1a), (2a), and (3a) generated the predicted rates given in Table 3.[13] Taking the predicted rates in this table as the rates that would have been paid in the years in question in the absence of interest rate regulation under the *Interest Rate Adjustment Act,* the losses to savers in terms of interest rates are the differences between the predicted rates and the actual rates paid.

TABLE 3

PREDICTED ($\hat{\imath}$) AND ACTUAL AVERAGE INTEREST RATES (IN PERCENT) PAID AT SAVINGS INSTITUTIONS

	$\hat{\imath}_{SL}$	i_{SL}	$\hat{\imath}_{CB}$	i_{CB}	$\hat{\imath}_{MS}$	i_{MS}
1968	4.88	4.68	4.46	4.25	4.88	4.76
1969	5.22	4.80	4.96	4.34	5.17	4.89
1970	5.61	5.06	5.49	4.72	5.53	5.01

By multiplying the average stocks of savings deposits at each

---

[12] The average maximum rates reported above are weighted averages of the maximum rates payable on the different classes of deposits with the weights being the fraction of total time and savings deposits (net of CDs) at commercial banks which were held as the given deposit class. The shortfall of the actual average rates in 1971 and 1972 from the average maximum rates payable in those years is consistent with the outcomes in 1968, 1969, and 1970, when the actual average rates paid were from 6 to 10 basis points less than the average maximum rates payable in those years.

[13] For example, the prediction for the savings and loan rate in 1970 is given by

$$\hat{\imath}_{SL}(1970) = 0.202 + 0.843\,\hat{\imath}_{SL}(1969) + 0.134\,i_{G}(1970).$$

savings institution (Table 4) by the appropriate interest rate loss, an estimate of the interest income lost in each year by savers at the three types of savings institutions was obtained. These results are reported in Table 5. The total estimated loss in interest income at all three savings institutions is $5.18 billion.[14]

TABLE 4

ANNUAL AVERAGE STOCK OF TIME AND SAVINGS DEPOSITS (IN $ BILLIONS) AT SAVINGS INSTITUTIONS

	TIME & SAVINGS DEPOSITS OTHER THAN LARGE CDs AT COMMERCIAL BANKS	SAVINGS DEPOSITS AT MUTUAL SAVINGS BANKS	SAVINGS CAPITAL AT SAVINGS AND LOAN ASSOCIATIONS
1968	171 2	62.5	127 5
1969	183 4	66 0	133 7
1970	191 3	68 9	139 3

TABLE 5

ESTIMATED INTEREST INCOME (IN $ BILLIONS) LOST BY SAVERS DUE TO INTEREST RATE REGULATION UNDER THE *INTEREST RATE ADJUSTMENT ACT* OF 1966

	ON SAVINGS CAPITAL AT SAVINGS AND LOAN ASSOCIATIONS	ON TIME AND SAVINGS DEPOSITS OTHER THAN LARGE CDs AT COMMERCIAL BANKS	ON SAVINGS DEPOSITS AT MUTUAL SAVINGS BANKS
1968	0 26	0 36	0 08
1969	0 56	1 14	0 18
1970	0 77	1 47	0 36
TOTAL	1 59	2 97	0 62

The estimate of interest income losses in excess of $5 billion during 1968 through 1970 does not include the continuing effects of interest rate regulation since 1970. There are two major effects. The first we have already noted in the discussion of commercial bank rates in 1971 and 1972, namely, that the maximum rates payable by commercial banks still hindered their ability to behave in the manner implied by the rate adjustment equation.[15] Second, the effective interest rate ceilings in the 1968–1970 period reduced the level from which savings and loan associations and mutual savings banks began the rate adjustment process when their ceilings were no longer effective.

---

[14] Since these estimates are based on the served level of deposits rather than on the level of deposits which would have prevailed at the forecast savings deposit rates, they are somewhat conservative in terms of the total interest foregone by potential savings depositors. However, if we assume that these potential depositors found alternative savings opportunities, the incremental loss which is neglected in my analysis would involve the difference in the forecast savings deposit rate and the rate which they obtained elsewhere (adjusted for risk and transaction costs). Even if we assume that the difference between the served level of deposits and the level which would have prevailed at the forecast rates represents foregone savings, the magnitude of the loss would involve the difference between the value of the foregone savings and the value of the added consumption. This point was brought to my attention by Paul MacAvoy.

[15] This effect alone resulted in additional opportunity losses on the order of $1.4 billion.

3. Some final observations on the interest income losses

The estimated income losses to savers at savings institutions due to interest rate regulation are important fractions of interest earnings during the years in question. The total interest income reported on individual income tax returns in 1968 was $16.8 billion; in 1969, $19.6 billion; and in 1970, $22 billion. The estimated losses on savings deposits are 4.2 percent of the total interest income reported in 1968; 9.6 percent in 1969; and 19 percent in 1970.

At least in the aggregate, savings institutions would have been able to sustain the added interest costs from net income over the period. For savings and loan associations, reserves and undivided profits increased by $2.82 billion over the period which was more than sufficient to cover the estimated cost of $1.59 billion. For commercial banks, net income for 1968 through 1970 totaled $10.2 billion as compared to the estimated added interest cost of $2.97 billion. For mutual savings banks, the change in general reserves was $0.74 billion compared to the estimated addition to interest cost of $0.62 billion.[16]

Finally, some evidence is available on the importance of benefits resulting from nonprice competition by savings institutions during the period. In the case of savings and loan associations, premiums and gifts given to savers are reported as advertising expenses. The ratio of advertising expenses to gross income at FSLIC-insured savings and loan associations trended downward from 1955 through 1967. Advertising expenses in 1968, 1969, and 1970 were in excess of those implied by an extrapolation of the trend. Even if the entire excess over the trend is taken as an added benefit to savers, the effect is to offset only $0.04 billion of the estimated losses at savings and loan associations.

In summary, the interest income in excess of $5 billion lost as a result of ceilings on interest payable under the *Interest Rate Adjustment Act of 1966* is an important fraction of total interest income for 1968 through 1970. The arguments brought to the defense of savings deposit rate regulation need to take these losses into account.

References

1. Kane, E. "Short-changing the Small Saver: Federal Discrimination against the Small Saver during the Vietnam War." *Journal of Money, Credit, and Banking*, Vol. 2, No. 4 (November 1970), pp. 513–522.
2. Malinvaud, E. *Statistical Methods of Econometrics*. Chicago: Rand McNally & Co., 1966.
3. Mullineaux, D. J "Deposit Rate Ceilings and Noncompetitive Bidding

---

[16] As Fischer Black has pointed out in private correspondence, higher rates would have resulted in higher deposit growth rates. The question then becomes whether reserves after the higher interest costs would have been sufficient to support these deposit levels. Without a measure of what constitutes sufficient reserves, there is no obvious answer to this question.

for U. S. Treasury Bills." *Journal of Money, Credit, and Banking,* Vol. 5, No. 1, part 1 (February 1973), pp. 201–212.

4. PYE, G. AND YOUNG, I. "The Effect of Deposit Rate Ceilings on Aggregate Income." *Journal of Finance,* Vol. 27, No. 5 (December 1972), pp. 1023–1031.
5. UNITED STATES SAVINGS AND LOAN LEAGUE. *1973 Savings and Loan Fact Book.* Chicago: United States Savings and Loan League, 1973.

The implications of regulation for induced technical change

V. Kerry Smith

Associate Professor of Economics
State University of New York, Binghamton

The purpose of this paper is to extend the Averch-Johnson model of the regulated firm so as to allow for the endogenous selection of factor augmenting technical change. A simple model is developed which avoids the questions of timing, capitalization, and internalization of innovations and focuses on the incentives to a firm for innovational choice under cost minimizing conditions versus several formulations of the regulated setting. The results suggest that regulation can distort the innovational selections. Moreover, in the case of the profit maximizing firm subject to "fair" return on investment regulation, the innovational choices will reinforce the static overcapitalization tendencies.

1. Introduction

■ One of the most influential contributions to the modeling of the behavior of the regulated firm was made in the development by Averch-Johnson[1] (hereafter A-J) of a static framework for analysis of resource allocation and output decisions. The A-J model suggests that regulation on the rate of return earned by a profit maximizing firm provides an incentive for overcapitalization. The efforts to extend and refine this basic model have increased apace.[2] There is now a variety of theoretical representations for the regulated firm, so that empirical analysis and testing are necessary to discern which structure provides the most reasonable description of the firm's behavior.[3] An important aspect of these tests will be an examination of the derived demand functions for factor inputs in the presence of regulation in order to detect whether or not incentives for overcapitalization are present. If these efforts are to be successful, yet another refinement to the A-J model must be considered.

The selection of technical changes by a regulated firm must be recognized as a decision that is made jointly with its factor hiring decisions and thus the chosen technology will influence the character of input demand functions. Failure to do so can result in mis-

---

Thanks are due Charles J. Cicchetti and John Jurewitz for comments on an earlier draft. I would also like to thank Paul W. MacAvoy and an anonymous referee for their constructive comments on several earlier drafts of this paper. The usual disclaimer applies.

[1] In [2].

[2] See Bailey and Coleman [4], Baumol and Klevorick [6], Klevorick [13], McNicol [15], Stein and Borts [22], Takayama [23], and Zajac [25] as examples of these developments.

[3] Robert M. Spann of Virginia Polytechnic Institute and State University has begun some of this empirical work in several working papers.

interpretation of any empirical findings. One such case is in Kahn's seminal work[4] on the economics of regulation. His discussion of John Kendrick's analysis of productivity trends from 1899 to 1953 for regulated utilities[5] indicates that he has implicitly extended the A-J model to form the following hypothesis:

> If the industry had a tendency to use excessively capital intensive methods of production . . . one would have expected it to show up in a more rapid relative rate of increase in labor productivity than in either capital or total factor productivity.[6]

Kahn then utilized Kendrick's evidence indicating that labor productivity had not increased greatly over capital or total factor productivity for regulated utilities to suggest that the tendencies described by the A-J model cannot be great. However, the A-J model does not allow for induced technical change and hence Kahn's interpretation of these results is inappropriate.

The purpose of this paper is to extend the A-J model to include the endogenous selection of technical change. The model developed in this paper is simple and avoids questions of the timing, capitalization, and internalization of innovations. Rather, the objective of the paper is to demonstrate that regulation can affect both the selection of factor inputs and the choice of technology by the firm. Section 2 reviews the Bailey-Malone framework,[7] which will be utilized to develop the model. The third section discusses the induced innovation model and alterations to it. The final section summarizes the paper.

2. The Bailey-Malone framework

■ Bailey and Malone have suggested that the regulated firm may have alternative objective functions including profit, return on investment, sales, and output. Moreover, the nature of the regulation can take one of several forms. After an examination of each combination of objective function and form for the regulatory constraint, they suggest that different selections can lead to different conclusions on factor hiring. In the analysis of induced technical change which follows, two of their four objective functions will be considered in conjunction with three forms of the regulatory constraint. The firm will be allowed to maximize either profit or sales, subject to a restriction on profit which is expressed: (1) as a "fair" return on investment, (2) as a fair mark-up on cost, or (3) as a fair profit per unit.[8] The models will maintain Bailey and Malone's notation and assume that the regulation is completely effective, so that the constraint equation is not an inequality.[9]

---

[4] In [9].

[5] In [12].

[6] Kahn [9], pp. 57–58. Given the present objection there are nonetheless good reasons for questioning his conjecture in the face of known deficiencies in productivity analyses such as Kendrick's. See Nadiri [16] for further discussion of these problems.

[7] In [5].

[8] The last cases, that of maximizing return on investment or output and the constraint of fixed profit, were omitted because their results were not sufficiently interesting to warrant inclusion. It should also be noted that Bailey and Malone provide a rather special interpretation of sales maximization. See Atkinson and Waverman [1] and Bailey [3].

[9] See Takayama [23].

Bailey and Malone define their symbols as follows:

$\pi = P(Q) \cdot Q(L, K) - WL - iK =$ profit
$r = [P(Q) \cdot Q(L, K) - WL]/K =$ return on investment
$Q(L, K) =$ level of output
$W =$ wage rate
$i =$ interest rate
$P(Q) =$ price of output (Q)
$s =$ regulatory approximation to firm's cost of capital (i)
$f =$ variable for statement of regulatory constraint. In the case of regulation calling for fixed return on capital f is defined as the difference between a fair rate of return and interest rate (i.e., $s - i = f$). Hence its definition will depend on the form of the constraint.
$L =$ labor
$K =$ capital.

Maximizing the predefined objective, subject to the constraints set forth in equations (18) through (22) of Bailey and Malone, yields first-order conditions which allow some interpretation of the resource allocation implications. If X is the objective function and $(X - Y)$ the constraint, both functions of labor and capital, then these first conditions can be written generally as in (1) through (3), with λ a Langrangian multiplier:

$$(1 - \lambda)X_L + Y_L = 0 \tag{1}$$
$$(1 - \lambda)X_K + Y_K = 0 \tag{2}$$
$$X - Y = 0. \tag{3}$$

Table 1 extracts from the Bailey-Malone Figure 3 the resource allocation ratios which equal the marginal rate of substitution between labor and capital for each of the six cases we shall examine. This model will now be altered so that the factor augmenting disembodied technical change is assumed to be chosen by the firm in conjunction with its selections of factor inputs.

TABLE 1

BAILEY-MALONE RESOURCE ALLOCATION RATIOS[a]

REGULATORY CONSTRAINT	MANAGEMENT OBJECTIVE	
	PROFIT	SALES
FAIR % INVESTMENT	$\frac{I}{W} - \frac{\lambda}{1-\lambda} \cdot \frac{S - I}{W}$	$\frac{S}{W}$
FAIR % COST	$\frac{I}{W}$	$\frac{I}{W}$
FAIR PROFIT PER UNIT	$\frac{I}{W}$	$\frac{I}{W}$

[a] THIS TABLE IS DERIVED FROM BAILEY AND MALONE [5], p 137. FOR A DISCUSSION OF ITS DERIVATION AND IMPLICATIONS SEE PP 136-42 OF THEIR MANUSCRIPT

3. A simple model of induced technical change

■ It is convenient to define a technology as consisting of the set of alternative procedures for combining materials and services in order to construct other commodities and services, conventionally designated as the outputs of the process.[10] This information is independent of political, social, and economic factors. Thus, at a given point in time the technology in existence will not necessarily be related to the economic or environmental conditions in that same time period.[11] However, over time this technology set will respond to economic and institutional incentives.[12]

---

[10] See Shephard [20], pp. 13–23.

Perhaps the only investigation of the implications of technical change for the regulated firm is a carefully developed study by Westfield,[13] which explores the interrelations between innovative behavior and the methods selected for regulation. The model which shall be outlined herein differs in an important respect from the model of Westfield. In his analysis, Westfield examines the effects of alternative predefined sets of innovations on the regulated firm. Consequently the designation of these types of technical change is made exogenously to the model. By way of contrast, a model of induced technical change allows the firm to select its innovations as part of its overall allocation decisions.[14] Such a model can address the question implicitly posed by Kahn—namely, what effect does regulation have on the technology selected by the firm?

Hicks[15] introduced the concept of induced technical change by postulating that innovations will respond to changes in the relative factor prices. While there has been a variety of questions raised regarding this model,[16] a framework which assumes that the firm selects innovations based on its anticipations of relative factor prices is now generally accepted. The present model remains in this tradition, though it does simplify the mechanism. The firm's production function is redefined in terms of augmented factors [i.e., $Q = Q(a_1L, a_2K)$]. The augmentation coefficients are assumed to be constrained by an innovational frontier with a concave shape. Equation (4) provides a general statement of the relationship:

$$T = g(a_1, a_2). \tag{4}$$

---

[11] Environmental conditions refer here to the socio-economic frame of reference rather than to the conditions of the ambient environment.

[12] See Ruttan [18].

[13] In [24].

[14] A more complete development of the induced technical change model is available in Kamien and Schwartz [10, 11].

[15] In [8].

[16] Hicks' [8] model ot induced innovation was at first interpreted as postulating that technology will respond to changes in the relative factor prices. Salter [19] subsequently argued against the Hicksian framework, noting that all inputs are equal in their contribution to costs at the margin. That is, under perfectly competitive factor and product market conditions, cost minimizing firms can be expected to hire their variable factor inputs until the ratio of each one's marginal physical product to price was equal to that of all others. This rule implies an equalization of the marginal costs associated with each input. Fellner [7] revitalized the Hicks model by suggesting that innovations are based on expected changes in relative prices. Since his work, there have been a number of further developments at both the macro and micro levels. It is probably safe to say that the basic philosophy of the induced innovations model for the firm has been accepted.

T measures the amount of innovative effort (e.g., the time of scientists) expended in developing the new means of combining factor inputs. The concave shape for the frontier is the result of diminishing returns in the process of research and development. Following Machlup,[17] one might suggest that scientists work efficiently by solving the easiest problems first. As more resources are devoted to research, the more difficult, costly, and frequently less significant problems are addressed under each activity type. Innovative effort is assumed to be priced at θ dollars per unit.

The firm can select both L and K and the levels of its augmentation coefficients (i.e., a_1 and a_2), subject to the constraints of a production function and an innovation frontier. The model, therefore, assumes that the firm is integrated backward in the development of its innovations. Moreover, technical change is not cumulative. When θT is zero the technology reverts to a base level.[18] While this approach seems quite restrictive, it should not distort our perceptions of the incentives present for the selection of innovations. Kamien and Schwartz's[19] work with models that postulate the discounted sum of profits for the firm's objective function indicates that the allocation of the research budget to minimize the cost of producing a given output will also yield the greatest increment to the defined profit function.[20] Hence, our static model will serve to discern the direction of technical change fostered by regulation.

Since the first-order conditions are essentially similar to those of Bailey and Malone with the addition of conditions for a_1, a_2, and our additional constraint, we shall report only the outcome of the first-order conditions. After some algebraic manipulation it can be shown that induced technical change requires the firm to equate the marginal rate of substitution between a_1 and a_2, measured along the innovational frontier, with the marginal rate of substitution between L and K weighted by the capital-labor ratio. For the cost minimizing firm without regulatory constraints this amounts to a finding which requires the relative augmentation coefficients to be proportional to the share of total costs accounted for by each factor, the proportionality constant being the marginal rate of substitution along the innovational frontier as we have in equation (5):

$$\frac{a_2}{a_1} = MRSI \cdot \frac{iK}{WL}$$

with

$$MRSI = -\frac{da_2}{da_1}. \tag{5}$$

---

[17] In [14].

[18] Assuming that the production function is homogeneous of degree one, then if $\theta T = 0$, we have:

$$Q = Q_L \cdot L + Q_K \cdot K.$$

Alternatively, with $\theta T \neq 0$, output is given by:

$$Q = a_1 Q_L \cdot L + a_2 Q_K \cdot K.$$

[19] In [10, 11].

[20] Kamien and Schwartz [10], pp. 6–10. In this case their model assumes a Cobb-Douglas production function. In [11] they develop a similar model with a somewhat more general production specification.

Consider the case of profit maximization with the first form of the constraint, then the relative augmentation coefficients will be given as in equation (6):

$$\frac{a_2}{a_1} = MRSI \cdot \frac{\left(1 - \lambda \frac{s}{i}\right)}{(1 - \lambda)} \cdot \frac{C_K}{C_L}, \tag{6}$$

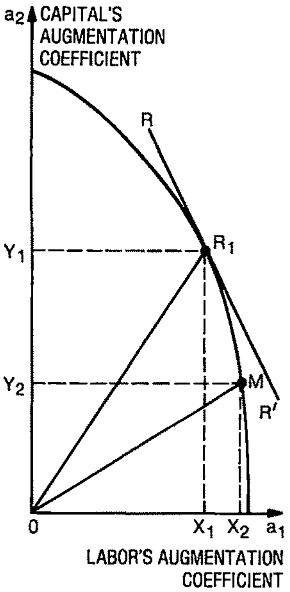
where

C_K = total cost of capital, and
C_L = total cost of labor.

Thus if $0 < \lambda < 1$, $s > i$, and $\lambda s < i$,[21] the firm will not only overcapitalize, but it will also select a technology which augments labor more than capital. In terms of Figure 1 if OX_1 is the value of a_1 and OY_1 is the value of a_2 under cost minimizing conditions, then the values corresponding to the selections of the regulated firm might be $OX_2 > OX_1$ and $OY_2 < OY_1$. This conclusion implies that the firm will select technologies which accentuate the allocation effects of regulation. Hence, there is reason to suspect that over time technology may serve to increase the discrepancies between the resource allocations selected by the regulated firm and those derived from a cost minimizing framework.

FIGURE 1

INNOVATIONAL CHOICE


a2
CAPITAL'S AUGMENTATION COEFFICIENT
R
Y1
R1
Y2
M
R'
0
X1
X2
a1
LABOR'S AUGMENTATION COEFFICIENT

In those cases where the allocation of resources between labor and capital is not affected by regulation, then the technology will also conform to that dictated by cost minimization. There remains, however, a case for which there is some distortion in resource allocation. With sales maximization, and the fair percentage return on investment, assuming $s > i$, the firm will tend to use an inefficient quantity of labor. As equation (7) indicates, this choice is again accentuated in the technology:

$$\frac{a_2}{a_1} = MRSI \frac{sK}{WL}. \tag{7}$$

Given $sK > C_K$ and this specification for the regulatory constraint, (a_2/a_1) will exceed the cost minimizing selection. The remaining cases conform to cost minimizing for the firm. Thus, given that our model reflects the behavior of the regulated firm, these findings indicate that technical change can serve to worsen the inefficiency in resource allocation predicted by the A-J model.[22]

---

[21] For a discussion of some of these considerations, see Baumol and Klevorick [6], pp. 166–167.

[22] It is possible to derive the same conclusion as equation (6) by using McNicol's [15] expense minimization framework. The present model might be respecified with the objective of minimizing E, subject to the rate of return constraint and augmented "labor requirements function" as follows:

$$E = iK + WL + \theta T + \lambda[R(Q) - sK - WL] \\ + \phi[a_1 L - h(Q, a_2 K)],$$

where all terms are defined as in the paper. $R(Q)$ is the total revenue function and $h(Q, a_2K)$ the augmented labor requirements function.
The first-order conditions are given as:

$$E_L = W - \lambda W + \phi a_1 = 0$$
$$E_K = i - \lambda s - \phi h_2 a_2 = 0$$
$$Ea_1 = \frac{\theta \partial T}{\partial a_1} + \phi L = 0$$
$$Ea_2 = \frac{\theta \partial T}{\partial a_2} - \phi h_2 K = 0,$$

It is reasonably straightforward to consider relaxing some of the assumptions of the previous models. One of the most important of these assumptions concerns the character of the innovation process. Our models have assumed that the firm is vertically integrated in the production of innovations. Suppose that the firm purchased the information on process reorganization in the market, then the augmentation coefficients, a_1 and a_2, would exchange for prices, say θ_1 and θ_2. If this technical information is available in perfectly competitive markets, so that the supply is perfectly elastic to the individual firm, then the result is equivalent to the case of a straight line (constant *MRSI*) innovation frontier, with $MRSI = \theta_1/\theta_2$. To the extent that markets for innovational information are not competitive, then the character of the supply functions will also serve to affect the choice of technology. In this case, *MRSI* is replaced by the ratio of the marginal augmentation costs as in equation (8):

$$MRSI = \frac{\theta_1\left(1 + \frac{1}{\epsilon_1}\right)}{\theta_2\left(1 + \frac{1}{\epsilon_2}\right)}, \tag{8}$$

where ϵ_1 and ϵ_2 are elasticities of supply for innovation (i.e., a_1 and a_2, respectively).

In order to determine the implications of these changes for the selection of technical change we need only substitute for *MRSI* from (8) into the appropriate equation. The net effect of two factors must be considered in determining the direction of technical change. Both the markets for the innovation (i.e., a_1 and a_2) and the effects of regulation will influence the firm's selection for technology. It should, however, be noted that our comparison was with the cost minimizing firm under similar conditions. Consequently, there remains the differential due to regulation when both are assumed to face the same kinds of markets for innovative inputs.

In order to argue that imperfections in the market for innovations will either accentuate or ameliorate the effects of regulation, it is necessary to demonstrate that these conditions are unique to the regulated firm. That is, we must be prepared to document differentials in the markets for innovations faced by regulated firms as opposed to the cost minimizing "benchmark" firm.

A further assumption in our analysis has been that the input markets and the markets for innovations are independent. Should there be a dependency between the two, namely that the suppliers also serve to promote innovations, then our model must be reformulated. Westfield has sketched the outlines of the results for such cases in two propositions. He notes that if input markets are competitive, then there is little incentive for the supplier to promote innovations in customers' production processes, since the

---

as well as the partial derivatives with respect to the Lagrangian multipliers. Solving for (a_2/a_1), we have equation (6). It should also be noted that including the costs of innovation in the regulatory constraint (i.e., reformulating $R(Q) = sK + WL + \theta T$) does not affect the relationship for (a_2/a_1) as in (6).

seller cannot extract any fraction of the rewards that will result. Under monopoly conditions the seller will contrive a scarcity of the innovative information in the same manner as he would for the inputs he sells.[23]

The models might be extended to consider the determination of the level of research expenditures made for each type of innovation. Following a basic model outlined by Nordhaus,[24] the present framework can be altered so that the firm selected its research expenditures to maximize profits over the life of the innovation (t years), subject to production and regulatory constraints in each period.[25] Research funds would be allocated to each activity (i.e., a_1 and a_2) until the marginal value product equaled the marginal cost. For the profit maximizing firm subject to the "fair" return on investment constraint the results conform to equation (6) with *MRSI* replaced by the ratio of the marginal products of research resources in the development of a_2 and a_1. Thus a rather simple structure permits a variety of extensions with little alteration in the conclusions of the analysis.

4. Conclusions and implications

■ The models developed in this paper indicate that regulation can affect not only resource allocation, but also the selection of technical innovations by the firm. In general, our findings suggest that the innovations will be selected to support the resource allocation decisions. Thus a profit maximizing firm subject to "fair" return on investment regulation will overcapitalize and select those technical changes which will allow it to continue to do so—namely labor augmenting innovations. While these models are greatly simplified, there is reason to believe that the tendencies they predict will hold for intertemporal specifications of the firm's objective function as well.

The policy implications of this analysis have several dimensions. First, it is generally argued that regulation on the allowed rate of

---

[23] Westfield [24], pp. 41–42.

[24] See [17], pp. 16–34.

[25] In this case our objective function might be respecified as:

$$\int_0^t \pi^*(x) e^{-rx} dx,$$

where r is the firm's rate of discount

$$\pi^*(x) = P(t) \cdot f[a_1(S_1)L, a_2(S_2)K] - W(t)L - i(t)K$$
$$- \theta(S_1 + S_2) - \lambda[R(Q) - W(t)L - sK].$$

S_1 and S_2 are allocations of innovative efforts made in the first period. $a_1(S_1)$ and $a_2(S_2)$ are augmentation functions, which are related to the allocation of research effort. Both are assumed to be subject to diminishing returns.

Integrating this objective function while holding wages, price of capital, and product prices constant yields an expression for the present value of the constrained profits. Applying conventional rules for maximization, we derive an expression similar to (6) where *MRSI* is replaced by

$$\frac{\dfrac{da_2}{dS_2}}{\dfrac{da_1}{dS_1}}\,.$$

return for utilities may impose costs on society in terms of inefficiency in resource utilization, but at any point in time these costs may be inconsequential. The inclusion of the inducement effects of regulation for technical change suggests that the costs of these inefficiencies may grow over time. That is, if the incentives for the choice of technical innovation are maintained over time, the path of technical changes that are selected will serve to support further overcapitalization. Second, on a more optimistic perspective, to the extent that the regulated utilities are not vertically integrated into research and innovation, the control of the pattern of external research and development may provide scope for ameliorating the resource misallocative effects of regulation. Moreover, in some cases this potential may be quantitatively important. For example, in the case of the electric power industry, technological change has occurred at a faster rate than in most other sectors of the economy. However, the R & D support has been undertaken almost exclusively by the equipment industry and the federal government, wherein such control could be easily exercised.

Finally, empirical testing of the variations on the A-J model will need to account for the effects of regulation on both factor input selections and innovation.

References

1. Atkinson, A. B. and Waverman, L. "Resource Allocation and the Regulated Firm: Comment." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 283–287.
2. Averch, H. and Johnson, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
3. Bailey, E. E. "Resource Allocation and the Regulated Firm: Comment on the Comments." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 288–292.
4. ——— and Coleman, R. D. "The Effects of Lagged Regulation in an Averch-Johnson Model." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 278–292.
5. Bailey, E. E. and Malone, J. C. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129–142.
6. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
7. Fellner, W. "Two Propositions in the Theory of Induced Innovations." *Economic Journal*, Vol. 71, No. 2 (June 1961), pp. 305–308.
8. Hicks, J. R. *The Theory of Wages*, 2d ed. New York: St. Martin's Press, 1968.
9. Kahn, A. E. *The Economics of Regulation*, Vol. 2. New York: John Wiley & Sons, 1971.
10. Kamien, M. I. and Schwartz, N. L. " 'Optimal' Induced Technical Change." *Econometrica*, Vol. 36, No. 1 (January 1968), pp. 1–17.
11. ———. "Theory of the Firm with Induced Technical Change." *Metroeconomica*, Vol. 23, No. 3 (September/December 1971), pp. 233–256.
12. Kendrick, J. W. *Productivity Trends in the United States*. Princeton: Princeton Univ. Press, 1961.
13. Klevorick, A. K. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 122–153.

14. Machlup, F. "The Supply of Inventors and Inventions," in R. R. Nelson, ed., *The Rate and Direction of Inventive Activity*, Princeton: Princeton Univ. Press, 1962.
15. McNicol, D. L. "The Comparative Statics Properties of the Theory of the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 428–453.
16. Nadiri, M. I. "Some Approaches to the Theory and Measurement of Total Factor Productivity." *Journal of Economic Literature*, Vol. 8, No. 4 (December 1970), pp. 1137–1177.
17. Nordhaus, W. D. *Invention, Growth, and Welfare*. Cambridge: The M.I.T. Press, 1969.
18. Ruttan, V. W. "Technology and the Environment." *American Journal of Agricultural Economics*, Vol. 53, No. 5 (December 1971), pp. 707–717.
19. Salter, W. E. G. *Productivity and Technical Change*. Cambridge: Cambridge Univ. Press, 1960.
20. Shephard, R. W. *Theory of Cost and Production Functions*. Princeton: Princeton Univ. Press, 1970.
21. Smith, V. K. "The Implications of Common Property Resources for Technical Change." *European Economic Review*, Vol. 3, No. 4 (December 1972), pp. 469–479.
22. Stein, J. L. and Borts, G. H. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 62, No. 5 (December 1972), pp. 964–970.
23. Takayama, A. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255–260.
24. Westfield, F. M. "Innovation and Monopoly Regulation," in W. M. Capron, ed., *Technological Change in Regulated Industries*, Washington, D. C.: The Brookings Institution, 1971.
25. Zajac, E. E. "A Simple Geometric Explanation of Averch and Johnson." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117–125.

The regulatory implications of three alternative econometric supply models of natural gas

Robert S. Pindyck

Assistant Professor of Economics
Sloan School of Management
Massachusetts Institute of Technology

Three different econometric models of natural gas exploration and discovery have appeared in recent issues of this Journal. By estimating all of them over the same time period and with the same data base, and then by simulating each of them as part of a complete supply-demand model, we can compare the regulatory policy implications of each formulation. The reestimated models are seen to have very different price elasticities from their original versions, indicating possible structural change in the industry over the past two decades. The policy implications, in terms of ceiling price increases necessary to eliminate the current shortage of gas, differ considerably among the formulations and indicate no consensus on how gas supplies are likely to respond to price increases.

1. Introduction

■ The shortage of natural gas in the United States has grown substantially in the last few years and is now at the level where it probably exceeds ten percent of total demands. Prior discovery of the lowest cost, highest volume reservoirs is not a sufficient explanation for this state of affairs, since rising resource costs alone would have resulted in relative price increases in natural gas and not in a large and systematic shortage. The fact is that the Federal Power Commission in the early 1960s set limits on field prices for new production approximately equal to the unregulated prices of the late 1950s, and then held these prices fixed for most of the decade.[1] The combination of rising resource costs, general inflation, and frozen gas prices provided strong incentives for increased demands but no incentives for increased discoveries and production.

The shortage (and this rationale for the shortage) has resulted in efforts to increase or even to deregulate field prices. It is ex-

---

The author greatly appreciates the help and advice of Paul W. MacAvoy, for providing both direction and criticism. The author also profited from the suggestions of J. Daniel Khazzoom. Finally, the work would not have been possible without the help of Krishna Challa, Kevin Lloyd, and Phillip Sussman, who assembled the data base and carried out all of the estimations and simulations This work was supported by the National Science Foundation, under Grant #GI-34936.

[1] See Breyer and MacAvoy [1].

pected that higher field prices would add to reserves and result in more production each year out of reserves; at the same time, demands for production would be dampened as users of boiler fuel replaced the more expensive gas with oil products or coal. But a tendency to clear markets of excess demand is not a sufficient reason for deregulation, and if higher prices resulted only in payments of rents to owners of the resource, then much of the current support for deregulation might vanish.

The 1973–1974 energy proposals of the Administration call for a series of substantial price increases, to be administered by one or the other agency, as regulation winds down over the rest of the decade.[2] Opposition in the Senate has proposed instead new (and in some cases more rigid) procedures for the Federal Power Commission to use in allowing price increases. Some believe that these procedures, resulting in price increases of only a few cents per Mcf greater than on new contracts signed this year, would be sufficient to clear the market of excess demand. The choice of the proper "interim" policy for the next few years depends on the amount of additional exploration, discovery, and production that is likely to follow from the Administration or Senate price increases.

Estimates of the response of reserve accumulations and production to price changes have been made by a number of sources, some on the basis of experience in the industry[3] and others on the basis of more formal modeling exercises.[4] Three of these modeling exercises have appeared in this *Journal*. Here we attempt to evaluate the estimates of these models in the context of their implications for regulatory policy.

The first model which we shall examine, constructed by MacAvoy and Pindyck, describes gas reserve exploration and discovery as well as wholesale demand and supply of production. The second and third models, by Khazzoom and by Erickson and Spann, respectively, describe only reserve exploration and discovery. In this paper we reestimate the Khazzoom and Erickson-Spann models by using the same data as for the exploration and discovery part of the MacAvoy-Pindyck model.[5] Each of these models for exploration and discovery will then be simulated as a module of the MacAvoy-Pindyck overall model. We shall thus produce three alternative sets of simulation results for natural gas reserves, production, and demand; the first will use the original exploration and discovery equations that were part of that model, the second will use Khazzoom's equations for exploration and discovery, and the third will use the Erickson-Spann formulation. Each of the three models for exploration and discovery has a different degree of structural detail. By comparing the three sets of simulations, we shall attempt to evaluate the models, and to assess the appropriateness of the degree of detail—or the extent to which structural theory is helpful in explaining and predicting gas reserves.

---

[2] See MacAvoy and Pindyck [4].

[3] In National Petroleum Council [5].

[4] See Khazzoom [3], Erickson and Spann [2], and MacAvoy and Pindyck [4].

[5] As we shall see, some minor respecification of those models is necessary in order to make the variables consistent with the data used in the MacAvoy-Pindyck model.

In Section 2 we describe the three alternative models, and present the reestimated forms of the Khazzoom and Erickson-Spann models. Section 3 presents the different simulation results based on the three alternative exploration and discovery formulations. Historical simulations are performed first (in order to compare the fit of the three models), and then forecasts are made of natural gas production demand and supply through 1980 so that the long-run implications of these alternative formulations can be compared.

2. Three alternative formulations of natural gas exploration and discovery

■ We shall review the three models only briefly here, as they have already been described in detail in this *Journal*, and shall concentrate more on the reestimations of the Khazzoom and Erickson-Spann models. As we shall see, reestimating a model over a different time period can provide information about the appropriateness of its specification.

☐ **The MacAvoy-Pindyck model.** This work is the result of a two-year effort by a group at the Massachusetts Institute of Technology to develop an econometric model of the natural gas industry for use in studying the effects of alternative regulatory policies on the wellhead price of gas.[6] The model is a set of simultaneous equations organized into functional modules. The exploration and discovery module begins with an equation for total exploratory wells drilled for any district in any year as a function of previous revenues, previous average cost, and average risk as measured by the variance of previous success rates. Next, two equations relate the average discovery size of pools and fields of nonassociated gas and the average discovery size of associated gas to past prices of gas and oil, past drilling costs, and the cumulative number of wells drilled. (The relationship of discovery size to cumulative wells drilled is expected to be negative because of depletion effects that occur over time in a reservoir region in which many wells are drilled.)

Separate equations are specified for nonassociated and associated extensions and revisions. Nonassociated and associated extensions depend on exploratory drilling activity as well as on previous nonassociated and associated new discoveries. Associated and nonassociated revisions, on the other hand, depend largely on the previous year's change in reserves. Finally, additions to reserves are given by total new discoveries plus extensions and revisions:

$$DR = NDNA + NDA + XN + XA + RN + RA. \quad (1)$$

This reserve accounting identity excludes, of course, losses and changes in underground storage.

---

[6] This model was developed under National Science Foundation Grant #GI-34936. An initial version of that model is described in detail in MacAvoy and Pindyck [4].

A second module determines the production of gas out of reserves by assuming exogenous prices equal to marginal costs and then by relating marginal costs to actual reserve levels and production levels. A third module consists of a set of regional wholesale demand equations that relate demand to the size of final retail or wholesale markets, wholesale prices of gas, and prices of alternative fuels. Wholesale and field markets are linked together by a set of pipeline price mark-up equations together with a regional input-output table that determines the distribution of gas from production regions to wholesale consumption regions.

The part of the model which is probably most open to controversy is the set of equations that explain reserve additions. Little is known about the exploration and discovery process—about the incentives for further exploration, the opportunities for significant finds (dependent on how much undiscovered gas remains under the ground in the United States) and the relation between *a priori* opportunities and *a posteriori* finds. This problem was approached by specifying equations that offered an intuitive explanation for drilling activity, and then by fitting these equations to the available data. The approach included the assumption that drilling can be divided into two modes of behavior, depending on whether it is done *extensively* or *intensively*. The first reaches out into new areas, with the supposed result that the ratio of successful wells declines but the size of find per successful well increases, while the second results in a higher success ratio but smaller finds. However, as a result of data aggregation it is impossible to determine *ex post* whether drilling has been done in the extensive or intensive mode.[7] Regression equations fitted *ex post* had a tendency to show increasing size of finds on higher gas prices for nonassociated gas, but decreasing size on associated gas.

The estimated versions of all of these equations are shown in the MacAvoy and Pindyck model.[8] District dummy variables were used in all of these equations, and the same district dummies will be used in the Khazzoom and Erickson-Spann models.

□ **The Khazzoom model.** This model, developed by Khazzoom for the Federal Power Commission,[9] uses two equations to relate new discoveries (*ND*) and extensions and revisions (*XR*) to a distributed lag of past gas ceiling prices (*PG*), oil prices (*PO*), and prices of natural gas liquids (*PL*). The model can be thought of as a set of reduced form equations (or equivalently a "black box"), because it does not specify explicitly any structural relationships between prices, drilling activity, and discoveries. This is not to fault the model, since it is not at all clear what a correct specification for a structural form model of exploration and discovery would be, and the reduced form specification of Khazzoom's model is appropriate to a variety of "reasonable" structural forms.

We reestimated Khazzoom's new discovery equation and his extensions and revisions equation over the same 18 production districts and for the years 1965 through 1969 as in the MacAvoy-Pindyck model. Since we had no data available for natural gas liquids, this variable had to be omitted from our version of the Khazzoom model. The district dummy variables used by Khazzoom were the same as those used in the MacAvoy-Pindyck model, and these include *DDA* for Louisiana South, *DDB* for the Permian producing region, and *DDC* for Kansas, Oklahoma, and Texas Railroad Commission districts, 1, 2, 3, 4, and 10.[10]

---

[7] See Subrahmanyam and Challa [6].

[8] [4], pp. 476–479.

[9] See Khazzoom [3].

A comparison of the original and reestimated versions of Khazzoom's equations are shown in Table 1A. Note that the price

TABLE 1

COMPARISON OF ORIGINAL AND REESTIMATED VERSIONS OF KHAZZOOM AND ERICKSON-SPANN

A KHAZZOOM

DEPENDENT VARIABLE	CONST	DDA	DDB	DDC	$(PG_{-1} + PG_{-2})/2$	$PO_{-1} + PO_{-2})/2$	PG	PO	ND_{-1}	XR_{-1}	R^2
ND (1)	–*	–	–*	–*	+*	–			+*		0 790
(2)	–	+	+	+	+	+			+*		0.895
XR (1)	–*	–*	–*	–*			+*	–	+*	+*	0 630
(2)	–	+*	+*	–			–	+	+*	+*	0 814

B ERICKSON-SPANN

DEPENDENT VARIABLE	CONST	DDA	DDB	DDC	log PGD	log POD	log SR_{-1}	log AFX_{-1}	log WRATIO	R^2
log WXT (1)	–	–*	+*	–	+	+*	–*	+*	+*	0.972
(2)	+*	+	–	–	–	+	–	–	+*	0 995
·log SR (1)	+	+	–	–*	+	–				0 802
(2)	–*	–	+*	+	+*	+*				0 274
log SIZE (1)	+	–	–*	–	+	–	–			0 599
(2)	+	+*	+*	+*	+*	+	–			0 268

SIGNS OF COEFFICIENTS FOR (1) ORIGINAL AND (2) REESTIMATED VERSIONS
*INDICATES SIGNIFICANCE AT 95% LEVEL.

---

[10] The resulting equations (with the *t*-statistics in parentheses) are:

$$
\begin{aligned}
ND = \underset{(-0.38)}{-42463} &+ \underset{(1.41)}{1.647 \times 10^5 DDA} + \underset{(1.14)}{7.862 \times 10^4 DDB} \\
&+ \underset{(0.92)}{3.401 \times 10^4 DDC} + \underset{(0.23)}{1086.3(PG_{-1} + PG_{-2})/2} \\
&+ \underset{(0.25)}{10844(PO_{-1} + PO_{-2})/2} + \underset{(11.99)}{0.685 ND_{-1}} \qquad (2)
\end{aligned}
$$

$R^2 = 0.895 \qquad S.E. = 1.23 \times 10^5 \qquad F = 100.5$

$$
\begin{aligned}
XR = \underset{(-0.03)}{-16252} &+ \underset{(2.283)}{2.209 \times 10^6 DDA} + \underset{(2.55)}{1.201 \times 10^6 DDB} \\
&- \underset{(1.06)}{1.653 \times 10^5 DDC} - \underset{(-0.04)}{946.0\, PG} + \underset{(0.17)}{3.209 \times 10^4 PO} + \underset{(2.77)}{0.645 ND_{t-1}} \\
&+ \underset{(2.72)}{0.325 XR_{t-1}} \qquad (3)
\end{aligned}
$$

$R^2 = 0.814 \qquad S.E. = 6\ 11 \times 10^5 \qquad F = 56.2.$

variables (*PG* and *PO*) are statistically insignificant in the reestimated equations, although they were significant in Khazzoom's original estimates. The reader might wonder why this change in fit occurs. We have, of course, estimated Khazzoom's equations over a somewhat shorter time period (1964 through 1969 versus 1961 through 1969), and one might suspect that by truncating the estimation period we have eliminated just those years (1961 through 1963) over which prices and discoveries had their greatest variance.

In Table 2 we show, for eight representative production districts around the country, new contract prices for the years 1952 to 1969, and new discoveries and extensions and revisions for 1956 to 1969 (data on these latter variables were not available before 1956). Note that between 1960 and 1969 discoveries dropped almost consistently in Louisiana South, Mississippi, Kansas, Texas 6, and Texas 10. New contract prices also dropped in those districts between 1960 and 1963, then rose between 1963 and 1969, so that the elimination of the 1961 to 1963 period (at least for those districts) would indeed reduce the positive correlation between price and discoveries. In Texas 3 and Texas 4 discoveries increased from 1960 to 1967 and then decreased from 1967 to 1969, while prices showed some decrease between 1960

TABLE 2

PRICE AND DISCOVERY DATA FOR 8 REPRESENTATIVE DISTRICTS

NEW CONTRACT PRICES (cents per mcf)

	1952	1956	1958	1960	1963	1967	1969
LA SOUTH	15 3	18 9	21 8	22 1	19 4	19 9	21 1
MISSISSIPPI	NA	19 4	22 1	19 1	16.6	18 2	20 5
KANSAS	9 8	12 9	12.9	14 8	14.0	15.4	15 6
PERMIAN	9.0	10 1	15 2	15.1	15 3	16 9	16 7
TEXAS 3	13 6	15 6	18.8	16 9	13 1	16 0	17 0
TEXAS 4	7.0	13 5	15 8	17 2	16 0	16 1	18 3
TEXAS 6	7 8	11 8	11 4	12.4	13.0	14 0	16 0
TEXAS 10	8.6	14.6	18 6	16 8	16 8	16 9	20 0

DISCOVERIES (x 10^5 mcf)

	1956	1958	1960	1963	1967	1969
LA SOUTH	14 950	18 030	21 870	17 330	16 9900	11.820
MISSISSIPPI	0.434	0 742	2 092	0 503	0.0557	0.922
KANSAS	1.943	0 269	1 556	0 854	0 4468	0.129
PERMIAN	5.309	7.384	3.786	3.937	3 9710	3.512
TEXAS 3	4 054	3 794	4 191	4 636	6 6260	3 775
TEXAS 4	9.557	8 704	6.731	9.896	12 1800	5 289
TEXAS 6	1 399	1 033	2,959	0.979	0.2300	0.288
TEXAS 10	0.943	1 993	0.731	0.555	1.1130	0.371

EXTENSIONS AND REVISIONS (x 10^5 mcf)

	1956	1958	1960	1963	1967	1969
LA SOUTH	28 530	34 190	40.420	54 920	64.820	29 830
MISSISSIPPI	−0 295	4.059	0.316	−1 050	0.901	0.510
KANSAS	16 440	1 489	1.281	1 249	2.077	4.878
PERMIAN	24 310	−14.850	6.346	12.090	39.240	−0.689
TEXAS 3	9.286	4.397	3.744	6.208	2.762	−5.629
TEXAS 4	29 590	3 255	6 556	4 183	6 759	−1.078
TEXAS 6	2.643	3.818	6.116	1.694	1 085	0 662
TEXAS 10	3 438	6 945	−5 361	0.498	12 160	0.392

SOURCE AGA AND FOSTER ASSOCIATES.

and 1963 and then increased thereafter. Therefore, eliminating the 1961 to 1963 years from the estimation period should for these two districts improve the positive correlation between price and discoveries. Of course, our regressions are also picking up cross-sectional variance between price and discoveries, but this does not change much from year to year. Putting this together, it would be reasonable to expect that decreasing prices in the majority of districts between 1960 and 1963 contributed to the significance of Khazzoom's price terms when his equations were estimated over the longer time period.

Another reason for the behavior of the price variables in Khazzoom's formulation is that both of his equations contain a lagged dependent variable. In the new discoveries equation this variable accounts for most of the explained variance, so that small changes in the sample can result in large changes in the coefficient estimates for other variables—in fact the equation represents little more than a simple first-order autoregressive model for a random process. If the errors are serially correlated (and there is no reason to believe that they are not), then his estimates of the price coefficients will also be inconsistent.

☐ **The Erickson-Spann model.** The third alternative model is that of Erickson and Spann,[11] which is more structural in nature. The first equation of that model relates wildcatting activity logarithmically to prices, previous success ratios, and other geographical variables. The second, also logarithmic, relates the success ratio to prices and geological variables, and the last two relate average oil discovery size and gas discovery size respectively to prices and geological variables. Extensions and revisions are not covered by the model, but new discoveries are explained with more structural detail than is the case in the MacAvoy-Pindyck model.

In order to make the structural format of Erickson and Spann's model consistent with our data base, some reformulation was necessary. For example, we respecified the drilling equation to include *all* exploratory wells drilled instead of just new pool and new field wildcats. The average depth of drilling variable thus also applies to all exploratory wells and not just to wildcats. Erickson and Spann use two variables relating to past drilling activity, the first of these is wildcats drilled by major companies in a given district in a given year as a percent of total U. S. wildcats drilled by all companies in that year, and the second is wildcats drilled by major companies in a given district in a given year as a percent of total U. S. wildcats drilled by *those companies* in that year. We did not have access to these data, and therefore used a single variable (*WRATIO*), which is the number of total exploratory wells drilled in a particular district in a particular year as a fraction of total exploratory wells drilled in the entire United States. Finally, the variables used by Erickson and Spann for Texas shut-down days were omitted from our version of their model since data for them were unavailable.

The restimated version of the Erickson and Spann wildcatting equation uses the same logarithmic form as Erickson and Spann do, and relates total exploratory wells drilled to deflated gas prices (*PGD*) and deflated oil prices (*POD*), previous success ratios (*SR*), average footage (*AFX*), and the activity variable (*WRATIO*) mentioned above. Next, a success ratio equation relates that variable to the deflated price of gas and the deflated price of oil, again in logarithmic form. Finally, an equation is estimated that relates average gas discovery size (for both non-associated and associated gas) to deflated oil and gas prices, and the previous period's success ratio.[12]

---

[11] In [2].

The fit of this model leaves something to be desired, particularly for the success ratio and size equations, but what is more startling is that the results, including the signs of several of the coefficients, differ considerably from Erickson and Spann's original results (see Table 1B). For example, one sees reversals in the signs of the coefficients of the price of gas and average footage in the wells equation, and in the price of gas and the price of oil in the success ratio equation. Most important, the price elasticities that result from our estimates of these equations are very different from those implied by the original estimates. Our price elasticity of gas discoveries, for example, is 2.36 (the sum of the elasticities of —0.033 for wildcat drilling, 0.570 for the success ratio, and 1.826 for the average size), as compared with 0.69 for Erickson and Spann's original estimates. Certainly simulations of this model are likely to show a much greater response to price simulation than was the case when the original estimates were used, but what is disturbing is that these elasticities changed so dramatically when the model was estimated over a different time period.

Erickson and Spann estimated their original equations over the period 1946 to 1959, while we have reestimated them over the period 1964 to 1969. As can be seen from Table 2, new contract prices in all districts showed much greater percentage increases from 1952 to 1958 than they did from 1960 to 1969.

---

[12] The reestimated equations are:

$$\log WXT = \underset{(45.83)}{9.018} + \underset{(0.01)}{1.577 \times 10^{-4}} DDA - \underset{(-0.09)}{0.0029} DDB$$

$$- \underset{(-1\ 13)}{0.0172} DDC \underset{(-0.98)}{-0.033} \log PGD + \underset{(1.02)}{0.065} \log POD - \underset{(-0.44)}{0.0067} \log SR_{t-1}$$

$$- \underset{(0.36)}{0.0061} \log AFX_{t-1} + \underset{(121.63)}{1.006} \log WRATIO \qquad (4)$$

$$R^2 = 0.995 \qquad S.E. = 0.063 \qquad F = 2497$$

$$\log SR = \underset{(-7.91)}{-4.869} - \underset{(-0.70)}{0.148} DDA + \underset{(2.67)}{0.554} DDB + \underset{(1.22)}{0.121} DDC$$

$$+ \underset{(2.51)}{0.570} \log PGD + \underset{(3.68)}{1.419} \log POD \qquad (5)$$

$$R^2 = 0.274 \qquad S.E. = 0.443 \qquad F = 6.93$$

$$\log SIZE = \underset{(0.28)}{0.804} + \underset{(3.61)}{2.693} DDA + \underset{(1.92)}{1.464} DDB + \underset{(4.43)}{1.564} DDC$$

$$+ \underset{(2.20)}{1.826} \log PGD + \underset{(0.24)}{0.348} \log POD - \underset{(-0.81)}{0.309} \log SR_{t-1} \qquad (6)$$

$$R^2 = 0.268 \qquad S.E. = 1.568 \qquad F = 5.57.$$

New discoveries, on the other hand, did not show such a dramatic percentage increase over the earlier period, and it is thus reasonable to expect that Erickson and Spann's model would show much smaller price elasticities when estimated over the earlier period. While that partially explains why the elasticities are greater when the model is estimated over the later time period, their model nonetheless seems to be considerably overestimating the price elasticity. The problem may be in their average size equation, which indicates (according to the estimated price elasticity of 1.826) that small price increases tend to result in very large increases in discovery size. The equation has a low R^2 and an F-statistic of only 5.6, so that most of the variance in average size is unexplained and the significance of the equation as a whole is questionable. Since Erickson and Spann's three equations are logarithmic, the price elasticities add across equations, and a spurious elasticity in the size equation becomes reflected in the overall price elasticity for discoveries.

3. Simulations of the three models

■ The three formulations for exploration and discovery estimated above can now be simulated as part of the complete supply-demand gas model in order to determine their alternative impacts on production, supply, demand, and excess demand. We begin by performing an historical simulation—a "backcast"—of each model over the period 1965 through 1971. Some explanation is needed for how these simulations are performed.

The Khazzoom model predicts both new discoveries and extensions and revisions, so that the two equations of that model can simply be substituted for the seven equations of the MacAvoy-Pindyck model that predict wells, discoveries, extensions, and revisions. The Erickson-Spann model, however, predicts only new discoveries, so that a method was needed to generate extensions and revisions. We chose to substitute the three Erickson-Spann exploration and discovery equations for the equations of the MacAvoy-Pindyck model that also predict new discoveries, but to retain the four extensions and revisions equations of the latter model.[13]

The results of these simulations are shown in Table 3, together with the root-mean-square (RMS) and mean simulation errors for each variable.[14] Note that over this seven-year time range the

---

[13] In performing the historical simulation we used the *actual* values for associated and nonassociated new discoveries as *inputs* to the extensions and revisions equations rather than the predicted values generated by the Erickson-Spann discovery equations. In a sense this gives some "benefit of the doubt" to the Erickson-Spann formulation, since it prevents accumulated errors in new discoveries from affecting extensions and revisions.

[14] It should be pointed out that simulated *demand* for production will be slightly different for each model formulation, even though the only differences among the models are in the exploration and discovery equations. The reason for this is that demand for production depends on the wholesale price (either for sales for resale or for mainline sales). The wholesale price in turn is based on a pipeline mark-up over an average "rolled-in" wellhead

TABLE 3

HISTORICAL SIMULATIONS

NEW DISCOVERIES (TRILLIONS OF CU. FT.)

	KHAZZOOM	E-S	M-P	ACTUAL
1965	4.927	3.538	5.535	6.300
1966	4.404	3.281	5.223	5.852
1967	3.702	3.139	4.832	4.993
1968	3.269	2.806	4.961	2.619
1969	2.865	2.813	5.120	3.384
1970	2.567	3.242	5.091	4.705
1971	2.327	4.496	5.036	4.487
MEAN ERROR	−1.182	−1.289	0.494	
RMS ERROR	1.491	1.697	1.192	

ADDITIONS TO RESERVES (TRILLIONS OF CU. FT.)

	KHAZZOOM	E-S	M-P	ACTUAL
1965	16.291	15.272	17.269	20.603
1966	14.464	12.344	15.724	18.455
1967	12.966	10.966	13.340	20.270
1968	11.793	9.998	12.821	11.561
1969	10.869	9.334	12.761	7.687
1970	10.086	9.626	12.647	10.177
1971	9.501	11.068	12.328	9.127
MEAN ERROR	−1.715	−2.767	−0.298	
RMS ERROR	3.773	4.831	4.108	

SUPPLY OF PRODUCTION (TRILLIONS OF CU. FT.)

	KHAZZOOM	E-S	M-P	ACTUAL
1965	18.499	18.373	18.404	16.095
1966	18.816	18.518	18.616	17.297
1967	19.082	18.637	18.817	18.258
1968	19.249	18.673	19.007	19.261
1969	19.392	18.711	19.276	20.657
1970	19.779	19.084	19.854	21.927
1971	20.634	20.234	20.917	22.137
MEAN ERROR	−0.026	−0.486	−0.106	
RMS ERROR	1.569	1.799	1.470	

MacAvoy-Pindyck formulation performs best in terms of new discoveries, with an RMS simulation error of 1.19×10^6 versus 1.49×10^6 for the Khazzoom formulation and 1.70×10^6 for the Erickson-Spann formulation. The Khazzoom formulation has the lowest RMS error for additions to reserves, since the aggregated extensions plus revisions equation of that model "tracks" the historical data more closely than do the disaggregated MacAvoy-Pindyck equations. It is questionable how meaningful this is, however, since (1) the Khazzoom extensions and revisions equation depends on new discoveries, which are being under-

---

price. The average "rolled-in" price is computed separately for each production region, and depends on the current period's production level in that region (since in computing rolled-in prices, new contract prices are weighted proportionately to new production). Production out of reserves, on the other hand, is itself a function of the average rolled-in price and year-end reserves, which means that the rolled-in price and production out of reserves are determined simultaneously, each one depending on the other and on year-end reserves. Thus, new reserves and new production, which differ among models, will lead to different prices at the field and wholesale level—and ultimately to slightly different demands at wholesale.

predicted—so that if Khazzoom's new discoveries tracked better, his extensions and revisions would track more poorly; and (2) the autoregressive component of Khazzoom's equation is helping it pick up trend, but what is allowing the model to pick up trend may prevent it from picking up turning points. What we are really after is a tool for analyzing the impacts of alternative regulatory price policies, and the autoregressive nature of Khazzoom's model is likely to prevent it from being useful for this purpose.

The Erickson-Spann formulation has the largest RMS error for additions to reserves; but this is entirely a result of its underpredicting new discoveries (since it is being simulated by use of the extensions and revisions equations of the MacAvoy-Pindyck model, with actual values of new discoveries as inputs to those equations). The Erickson-Spann model also has the largest underprediction of the supply of production out of reserves, again because it underpredicts new discoveries. Finally, the simulated values for the demand for production out of reserves is roughly the same in all three formulations; again the differences are attributable to the different predicted rolled-in prices that result from different supply predictions, but these differences are not substantial on this side of the model.

It is also informative to simulate the three formulations over a future time horizon in order to compare their long-run dynamic behavior. We performed forecast simulations over the time period 1971 through 1980, using two alternative sets of assumptions about regulatory policy for each model simulated. The first policy, called "cost of service" regulation, represents historical average cost pricing by the FPC, which is taken to imply wellhead price increases of 1 cent per Mcf per annum on new contracts. Next, a "deregulation" policy is simulated in which new contract prices are allowed to rise by 15 cents in 1974 and then an additional 4 cents per annum after 1974. (For a detailed description of these alternative policies, see MacAvoy and Pindyck.[15]) In all of the forecasts, exogenous variables (such as the price of oil, per capita *GNP*) were assumed to follow the "medium" growth paths as described in that article.

The results of these forecasts are shown for additions to reserves and production in Table 4, and are not surprising. Forecasts using the Khazzoom model show a lack of price sensitivity, as there is almost no response in reserve additions to increases in the wellhead price. Thus, even under the "deregulation" policy, excess demand grows steadily and reaches a level of more than seven trillion cubic feet by 1980.

Forecast results using the Erickson-Spann exploration and discovery equations indicate that their model is extremely sensitive to price, so that under "deregulation" new discoveries increase about tenfold between 1971 and 1980, production out of reserves more than doubles in that time period, and we have excess *supplies* of about 18 trillion cubic feet by 1980.

Again, this is because of very large supply elasticities for new discoveries with respect to oil and gas prices in the Erickson-Spann

---

[15] In [4].

TABLE 4

FORECASTS OF THE THREE MODELS

ADDITIONS TO RESERVES (TRILLIONS OF CU FT)

	MACAVOY-PINDYCK		KHAZZOOM		ERICKSON-SPANN	
	(1)	(2)	(1)	(2)	(1)	(2)
1975	18 51	19 03	11 36	11 22	27 87	58 89
1976	21 89	24.09	11.57	11 57	27 79	65 30
1977	24 13	28 79	11.73	11 91	27 05	69 71
1978	25 99	32 57	11 86	12 21	26 28	73.65
1979	27 99	36.26	11 96	12 48	25 56	77 29
1980	30 27	40 26	12 04	12 72	24 91	80 64

PRODUCTION (TRILLIONS OF CU FT)

	MAC AVOY-PINDYCK		KHAZZOOM		ERICKSON-SPANN	
	(1)	(2)	(1)	(2)	(1)	(2)
1975	24 70	27 22	23 34	25 86	27 23	34 45
1976	25 83	28 80	23 59	26 35	28 31	38 35
1977	27 00	30 43	23 71	26 71	29 09	41 79
1978	28 20	32 08	23 82	27 03	29 64	44 91
1979	29 48	33 82	23 94	27 31	30 06	47 82
1980	30 87	35 70	24 09	27 59	30.38	50 55

(1) "COST OF SERVICE" POLICY
(2) "DEREGULATION"

formulation. Adding the supply elasticities of wells drilled, success ratios, and size, we find that the elasticity of total discoveries with respect to the deflated gas price is equal to about 2.36, and the elasticity of total discoveries with respect to the deflated oil price is equal to about 1.83. These elasticities apply only to new discoveries and not to total additions to reserve; the elasticities for additions to reserves are about twice as large, since new discoveries contribute to extensions and revisions. Thus a doubling of the wellhead price (under the "deregulation" policy) can result in a tenfold increase in discoveries.

The MacAvoy-Pindyck model shows excess demand of about ten trillion cubic feet in 1980 under "cost of service" regulation, and the elimination of excess demand by 1979 under the "deregulation" policy. Remember that what we are calling "deregulation" is not really deregulation—we are not permitting complete decontrol of the wellhead price, but simply allowing price to increase by a specified amount. We picked a price increase of 15 cents per Mcf in 1974 (with further price increases of 4 cents per annum), because that came close to doubling the wholesale price by the end of the decade (and as such would be "in line" with recent policies of the Cost of Living Council). Now the question, of course, is whether that price increase is too small (as suggested by the Khazzoom formulation) or is too large (as suggested by the Erickson-Spann formulation). In order to have markets clearing by 1980 with the Khazzoom formulation, it would be necessary to have a price increase in 1974 of something closer to 50 cents per Mcf. In order to have markets just clear by 1980 (with little or no excess supply) with the Erickson-Spann formulation, an additional price increase in 1974 of 2 or 3 cents (beyond the 4 cent per annum price increase) would probably be sufficient.

4. Conclusion

■ Policy proposals for dealing with the natural gas shortage have been seen to be very sensitive to the particular model that is used for exploration and discovery. The three alternative formulations for exploration and discovery that we have examined in this paper suggest very different price policies for dealing with the shortage (although this author's own judgment would favor the 15 cent price increase as being the most reasonable). Furthermore, when two of the models were reestimated over a different time period, their coefficient values changed considerably. This may be due in part to structural change in the industry, but it may also be due simply to an incorrect specification of the models.

The three models that have been examined here are probably representative of the current state of the art of econometric modeling of the natural gas industry, but they provide no consensus on how gas supplies are likely to respond to ceiling price increases. It is clear that a knowledge of the dynamic response of exploration and discovery to changes in the price incentive is crucial to the design of regulatory policy; unfortunately, it represents an area that is still not well understood.

References

1. Breyer, S. G. and MacAvoy, P. W. "The Natural Gas Shortage and the Regulation of Natural Gas Producers." *Harvard Law Review*, Vol. 86, No. 6 (April 1973), pp. 941–987.
2. Erickson, E. and Spann, R. "Supply Response in a Regulated Industry: The Case of Natural Gas." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 94–121
3. Khazzoom, J. D. "The FPC Staff's Econometric Model of Natural Gas Supply in the United States." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 51–93.
4. MacAvoy, P. W. and Pindyck, R. S. "Alternative Regulatory Policies for Dealing with the Natural Gas Shortage." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 454–498.
5. National Petroleum Council. *United States Energy Outlook*. Washington, D. C., December, 1972.
6. Subrahmanyam, M. and Challa, K. "The Oil and Natural Gas Exploratory Process in the United States: A Framework for Analysis." Unpublished research memorandum. October, 1973.

Determinants of research and development activity by electric utilities

Ronald P. Wilder

Assistant Professor of Economics
College of Business Administration
University of South Carolina

and

Stanley R. Stansell

Assistant Professor of Finance and Management Science
College of Business Administration
University of South Carolina

This paper develops a model of the determinants of R & D outlays of privately owned electric utilities and tests the model empirically with data for the years 1968 through 1970. The findings are that R & D outlays have an elasticity greater than one with respect to firm size, and are positively associated but relatively inelastic with respect to profitability. These results suggest that increasing firm size, either through merger or internal growth would have a favorable effect on R & D outlays. The results also imply that an increase in the regulated rate of return would not be an effective way to increase utility R & D outlays. There is no significant difference between the R & D behavior of utilities which do not fall under state regulation and those which are regulated.

1. Introduction

■ The question of whether large absolute or relative firm size is a necessary condition for the research and development outlays conducive to rapid innovation and technological progress has received considerable attention from students of the industrial structure of the manufacturing sector.[1] Research and development in the utility sector, on the other hand, have been the subject of little empirical research.[2] This paucity of research is rather surprising, given the availability of data and the renewed interest in the behavior of regulated firms in recent years.

The objective of the present study is to investigate the determinants of R & D activity in the electric utility sector. We construct and empirically test a model to explain interfirm differences in R & D outlays by using data for the period 1968 through 1970. We are particularly interested in the relationships between R & D outlays and firm size, profitability and the extent to which the utility is regulated. The large sample size and consistency of the data lend themselves to the study of these relationships.

---

[1] For a survey, see Markham [5].

[2] For a treatment of some aspects of the subject, see Capron [1].

2. Public utility research and development—the model

Rate regulation may be expected to have some effect on the firm's investment behavior, as is recognized by the Averch-Johnson hypothesis. Research and development activity, as a special type of investment, may also be influenced by the effect of regulation on the profit-possibility distribution the firm faces. As pointed out by Capron, regulation is likely to cut off both the upper and lower ends of this distribution, so that the firm is both protected against the risk of large losses and denied above normal profits which could result from innovative activity.[3] Because of this effect of the profit constraint on the Schumpeterian incentive to innovate, we have no *a priori* expectation of the effect of profitability on R & D activities in cases where the firm is regulated in this fashion.

Research by Hughes into the adoption and diffusion of new technology in the electric power industry finds that the most important technological changes have been associated with increases in scale of the generating unit and that the largest power systems have been the technological pioneers.[4] We employ total utility operating revenues as a size proxy to examine this relationship. It should be apparent that we are *not* testing hypotheses regarding economies of scale in research and development, but are dealing only with the determinants of R & D inputs as measured by annual outlays. Fisher and Temin have pointed out that studies examining R & D inputs have little relevance to the question of returns to scale in R & D.[5]

The ratio of electric to total operating revenue defines the degree to which the firm operates in the electric utility sector. We employ this variable to differentiate electric utility operations from other types of public utility activities. Revenues derived from other activities (such as gas distribution) presumably involve essentially static technologies. Our *a priori* expectation is that this variable should be positively correlated with R & D activity.

The degree of regulation variable is included in the equation to test the effect of state regulation on research and development outlays. Previous discussions of the effect of regulation, as noted above, have emphasized that regulation's effect on innovative activity could work in either direction because of restrictions on both ends of the profit-possibility distribution. Another plausible hypothesis is that some research and development activity may be an organizational slack item, and that regulation might lead to "X"-inefficiency within the regulated firm.[6]

The model which we test is specified as follows:

$$R = A_0 S^{A_1} E^{A_2} \Pi^{A_3} U^{A_4}, \qquad (1)$$

where

R is annual R & D outlay, in thousands of dollars;
A_0 is a constant term;
S is total utility operating revenues, in thousands of dollars;
E is the ratio of electric to total operating revenue;

---

[3] In [1], pp. 8–9.

[4] See [3].

[5] In [2].

[6] See Leibenstein [4].

Π is the rate of return on common equity, expressed as a percentage; and

U is the degree of regulation dummy variable, taking on the value 0 for a utility which operates primarily in a state in which electric utilities are state regulated and 1 if state regulation is not practiced.

The variables included in equation (1) are relatively self-explanatory, except for the profitability and regulation variables. The rate of return on common equity is defined as the ratio of earnings available for common stock to the average of the beginning and year-end balances in proprietary capital less preferred stock. As we have pointed out, the regulation dummy variable is based on whether or not a given utility operates primarily in a state which practices rate regulation of electric utilities. States in which electric utilities are not regulated include Minnesota, Nebraska, South Dakota, and Texas.[7]

3. Utility research and development outlays—the data

■ The source for all data used in this study is the Federal Power Commission.[8] The chief advantage of this source is the relatively high degree of uniformity among firms, since accounting conventions are specified by the FPC. Our sample contains all utilities reported in the FPC source, except for a small number for which zero values are reported for total operating revenues. Sample sizes are in excess of two hundred firms for each of the three years (209, 204 and 207 for the years 1968, 1969, and 1970). There are fifteen utilities in our sample which operate primarily in non-regulated states.

4. The estimating equation and empirical results

■ The estimating equation used is the logarithmic version of equation (1):

$$\log R = a_0 + a_1 \log S + a_2 \log E + a_3 \log \Pi + a_4 \log U + v, \quad (2)$$

where the a_i are the estimators of the A_i in (1) and v is an error term. Natural logarithms are used throughout so that the resulting coefficients are interpreted as elasticities.[9]

The logarithmic form enables us to look at the effect of relative size rather than absolute size. In addition, we can test directly whether or not research and development intensity (the ratio of R & D to operating revenue) increases with firm size as measured by total operating revenues by determining whether a_1 is significantly greater than one.[10]

---

[7] See [6]. Municipalities have franchise regulatory power in the non-regulated states. Electric utilities in Nebraska are publicly owned and therefore not included in our sample.

[8] See [7].

[9] To facilitate use of the double-logarithmic form, all variables except the dummy variable were coded by adding the constant one to each observation before taking logarithms. The dummy variable is considered to be 0–1 after logarithms have been taken.

[10] We also tried alternative estimating equations to test the sensitivity of the results to a form which permitted a nonzero intercept. We specified

The results of applying ordinary least squares regression analysis to equation (2) are shown in Table 1. The estimated coefficient for the size variable (a_1) is positive and significantly greater than both zero and one for each of the three years. We may therefore characterize R & D outlays as being relatively elastic with respect to total operating revenues, which implies that R & D intensity increases with firm size.

TABLE 1

REGRESSION COEFFICIENTS, RESEARCH AND DEVELOPMENT OUTLAYS RELATED TO TOTAL OPERATING REVENUE AND OTHER VARIABLES, ELECTRIC UTILITIES, 1968-1970

	CONSTANT	TOTAL OPERATING REVENUE	ELECTRIC/TOTAL OPERATING REVENUE	RATE OF RETURN	REGULATION DUMMY VARIABLE	COEFFICIENT OF DETERMINATION
YEAR	a_0	a_1	a_2	a_3	a_4	R^2
1968 (1)	−16.86	1.35	1.15	0.24	−0.13	0.74
(2)		(20.44)*	(4.17)*	(0.95)	(−0.38)	
(3)		(5.35)*	(0.56)			
1969 (1)	−17.30	1.29	1.14	0.72	−0.24	0.77
(2)		(20.64)*	(5.03)*	(3.07)*	(−0.77)	
(3)		(4.69)*	(0.64)			
1970 (1)	−18.30	1.40	1.20	0.47	−0.06	0.80
(2)		(24.06)*	(4.95)*	(2.12)*	(−0.18)	
(3)		(6.93)*	(0.83)			

NOTES ROW 1 REPORTS REGRESSION COEFFICIENTS AND R^2 (UNADJUSTED)
ROW 2 REPORTS t-RATIOS FOR NULL HYPOTHESIS $A_I = 0$
ROW 3 REPORTS t-RATIOS FOR NULL HYPOTHESIS $A_I = 1$
*ASTERISK DENOTES STATISTICAL SIGNIFICANCE AT THE 0.05 LEVEL (ONE-TAILED TEST)

The estimated coefficient for the relative importance of electric operations (a_2) is positive and significantly greater than zero for each of the three years. Although the estimated coefficient is greater than one for each year, the differences from one are not statistically significant. These results suggest that utilities perform more research and development associated with electric operations than with other (e.g., gas) operations.

The estimated coefficient for rate of return is positive and significantly greater than zero in two of the three years (1969 and 1970). The responsiveness of research and development to profitability is relatively inelastic, since the coefficients are all less than one.

The estimated coefficient for the regulation dummy variable has a negative sign for each of the three years, but is relatively weak and not significantly less than zero. These results suggest that the net effect of regulation on research and development outlays is virtually neutral, since there are only fifteen nonregulated firms in the sample.

---

both a quadratic form and a logarithmic form with an intercept term, but in neither case was the intercept term statistically significant. We are grateful to J. Michael Ryan for his help with this phase of the study.

5. Discussion and conclusions

■ Our findings that electric utility total research and development outlays are relatively elastic with respect to firm size are consistent with previous research findings that the largest utilities tend to be technological leaders. These results are quite stable over the three years studied, and, since 1970 is a recession year while 1968 and 1969 are years of relatively rapid economic growth, we believe that the results are not sensitive to the business cycle. The results suggest that increasing firm size, either through a merger or internal growth, would have a favorable effect on R & D outlays.

Several studies of nonregulated manufacturing firms suggest that, up to a certain firm size, R & D activities increase more than proportionally; beyond that size such activities generally do not increase proportionally and may in fact decline proportionally.[11] The threshold size of firm appears to vary from industry to industry. We have not found a similar threshold size in electric utilities.

The results also imply that an increase in the regulated rate of return would not be an effective way to increase utility R & D outlays (with an elasticity of R & D outlays to rate of return in the region of 0.5, a 1-percent increase in total net income in the utility sector—$3 million—would be associated with a 0.5-percent increase in R & D outlays, or only about $200,000, *ceteris paribus*).

The neutrality of R & D behavior with respect to regulatory status is interesting, but subject to more than one interpretation. Since regulation presumably affects the profit possibility distribution faced by the utility, we had expected the effect of regulatory status to be significant, although theoretically its effect could be either positive or negative. The finding that regulatory status has no significant effect could be interpreted to mean that state utility regulation is generally ineffective or that R & D outlays by most utilities represent a corporate consumption item.

References

1. Capron, W. M., ed. *Technological Change in Regulated Industries.* Washington, D. C.: The Brookings Institution, 1971.
2. Fisher, F. and Temin, P. "Returns to Scale in Research and Development: What Does the Schumpeterian Hypothesis Imply?" *Journal of Political Economy*, Vol. 81, No. 1 (January/February 1973), pp. 56–70.
3. Hughes, W. R. "Scale Frontiers in Electric Power," in [1].
4. Leibenstein, H. "Allocative Efficiency vs. '*X*-Efficiency.'" *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 392–415.
5. Markham, J. "Market Structure, Business Conduct, and Innovation." *The American Economic Review*, Vol. 55, No. 2 (May 1965), pp. 323–332.
6. U. S. Federal Power Commission. *Federal and State Commission Jurisdiction and Regulation.* Washington, D. C.: U. S. Government Printing Office, 1967.
7. ———. *Statistics of Privately Owned Electric Utilities in the United States, Annual.* Washington, D. C.: U. S. Government Printing Office.

---

[11] See Markham [5].

Dynamic aspects of public goods: a further analysis of the telephone system

Burckhard von Rabenau

Assistant Professor of City and Regional Planning
Ohio State University

and

Konrad Stahl

Doctoral Candidate
University of California, Berkeley

This papers reconsiders a dynamic model of the telephone system developed by Artle and Averous, which for the first time introduces public goods in a dynamic framework. With a more restrictive set of assumptions, their proposition showing self-sustaining growth in the demand for telephones is restated and discussed in detail. Comparative statics of the model are discussed to illuminate further properties which are due to the public good characteristics of the telephone system. Alternative interpretations of the dynamics of the model are given. Finally, a paradigm of urban growth is given to suggest applications of this type of model to other areas of economic analysis.

1. Introduction

■ In a recent article in this journal, Artle and Averous (henceforth referred to as AA) analyze static and dynamic aspects of commodities with private and public good properties.[1] Using the telephone system as example, they show in a static analysis that the telephone system, through its provision of access and through savings in transaction costs, possesses properties of a public good, and they provide in a dynamic analysis a theoretical explanation for the growth in the demand for telephone services over time. More specifically, they suggest that "a growth process in the demand for telephones can be explained jointly by the distribution of incomes and the cumulative public good property inherent in the telephone system."[2]

In our view, this paper provides several interesting contributions, first by introducing into formal analysis decisions among lumpy alternatives, second by providing a framework for the dis-

---

The authors have benefited from the Urban Economic Systems Seminar in the Department of Electrical Engineering, UCB, especially from discussions with Roland Artle and Pravin Varaiya, and also from a referee's comments. Both authors were supported by the Electronics Research Laboratory, University of California, Berkeley, through National Science Foundation Grant NSF-GK-10656x2.

[1] In [4].

[2] [4], p. 99.

cussion of growth processes that are peculiar to systems involving public goods of a cumulative nature.

Using more restrictive assumptions, we wish to illuminate in the present paper further properties of their dynamic model.[3] We first state and comment on the assumptions introduced and the proposition advanced and proven by AA and then describe the goals of our own analysis. Following AA, we introduce the following concepts:

Consider a population consisting of h individuals. Let y be the level of an individual's income, and $\Psi(y)$ a cumulative income distribution expressing the share of the total population with income below y. Let $x(t)$ be the number of telephone connections established up to time t. We disregard any "decay" of telephone connections. Furthermore, let p_x denote the subscription price of the telephone service, per unit of time. Any output demanded at time t is supposed to be instantaneously forthcoming.

The following assumptions are introduced by AA:

(A1) The total population h, the distribution of income $\Psi(y)$, and the price of telephone subscription are all independent of time and thus stationary.

(A2) The distribution of telephones among individuals is such that the order of new subscribers entering the telephone system is the order of their income:

$$x(t) = h\{1 - \Psi[z(t)]\}.$$

Here, $z(t)$ is the income level beyond which all individuals have access to the telephone system in period t.

(A3) The preferences of individuals can be represented by a utility function $U(c, \delta x)$, continuous and strictly increasing in c and δx, and where c is the consumption of private goods by the individual, x is the number of telephones to which the individual has access and $\delta = \begin{cases} 1 \\ 0 \end{cases}$ if the individual has $\begin{cases} \text{a} \\ \text{no} \end{cases}$ telephone.

(A4) At any time each individual maximizes his utility subject to his budget constraint:

$$\max_{c(t+1),\delta} U[c(t+1), \delta x(t)] \text{ subject to } c(t+1) + \delta p_x = y(t+1),$$

where the price of the private goods has been set at unity.

(A5) $x(0)$ and $x(1)$ are given exogenously, such that $x(1) > x(0)$.

The proposition advanced by AA is as follows: "Given (A1)–(A5), the sequence $\{x(t)\}$ increases." Our comments on this formulation are as follows:

(1) The proposition is correct only if (A5) refers to an endogenous change in telephone stock. Then it states that if the telephone stock was observed to grow in the previous period, then it will grow in the future. Thus the proposition does not show the existence of growth under some conditions but only its continuation.[4] We attempt to generalize this proposition. In particular, we

---

[3] [4], pp. 97–99.

[4] See also the comments after the proof of Proposition 1 below.

shall show that the telephone system grows or declines independently of any initial exogenous increase in telephone connections and that the growth is dependent only on the initial stock of telephones.

(2) Some assumptions, such as (A4), are concerned with individual behavior, while others are concerned with aggregative behavior, for instance (A2). We prefer to base the discussion entirely on assumptions about individual preferences and decision making. In replacing (A2) we therefore introduce restrictions on individual preference functions, such that the desired aggregative behavior is generated by individual decision making. This formulation enables us to analyze system behavior more extensively. Introducing, for convenience, differentiability of the preference functions, we show in a comparative static analysis further properties due to the public good characteristics inherent in the telephone system. In particular, it is shown that an expansion of demand can be accompanied by a price increase; that there may exist a multiplicity of equilibria; that an initial threshold has to be overcome in order to bring the growth process to a start; also, that if the good is produced at all, then there exists a unique saturation level of the market; and finally, that population and income changes result in rather unexpected changes in threshold and saturation levels. In addition, the graphical explanation given should contribute to a deeper understanding of system behavior.

(3) The periodicity introduced by AA remains unexplained. In particular, no information is supplied as to how the growth process varies with the length of the time period chosen. Alternative dynamics are discussed in Section 4.

(4) We finally suggest a simple reinterpretation of the model and its properties developed by referring to some aspects of urban growth.

2. Discrete time growth model of demand

■ In this Section, we shall introduce our modification and extension of AA's model. We continue to use the notation introduced above and the definition of the individuals' preference function given in (A3).[5] We also use the relationship introduced by AA[6] stating the number of telephones which must be in the system so that an individual with income y will be indifferent between owning and not owning a telephone. Using $w(y;p_x)$ to denote this number,

$$w(y;p_x) = \{x\colon U(y - p_x, x) = U(y, 0)\}. \qquad (1)$$

We may drop p_x as an argument when considering w for given p_x only. It follows that for $x > w(y)$ an individual will subscribe to a telephone, and for $x < w(y)$ he will not.

---

[5] This specification implies the simplifying assumption that the individual's consumption of telephone services is proportional to the number of telephone connections. Note also the simplification that the individual can gain access to the system only by subscribing to a phone. It should also be mentioned that throughout the analysis we disregard congestion phenomena in communication.

[6] In [4], footnote 11, p. 99.

Our assumptions are as follows:

(B1) $\Psi(y)$ is continuous and strictly increasing for $\underline{y} \leqslant y \leqslant \bar{y}$ where $\underline{y}$ and $\bar{y}$ are lower and upper income bounds with $0 \leqslant \underline{y} \leqslant p_x \leqslant \bar{y} \leqslant \infty$. Thus, there are no individuals with income $y < \underline{y}$ and $y > \bar{y}$, so $\Psi(y) = 0$ for $y \leqslant \underline{y}$ and $\Psi(y) = 1$ for $y \geqslant \bar{y}$. Also assume that $\Psi(y)$, h, and p_x are stationary.

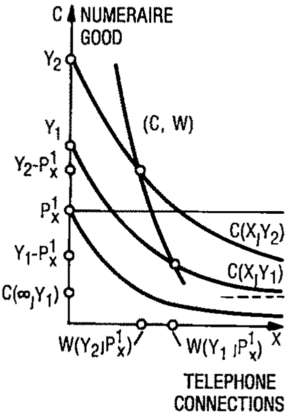
We shall be interested mostly in the number of individuals with income greater than y, $N(y)$, given by

$$N(y) = h[1 - \Psi(y)], \tag{2}$$

which is strictly decreasing for $\underline{y} \leqslant y \leqslant \bar{y}$ and constant elsewhere.

FIGURE 1a

CONSUMER'S INDIFFERENCE MAPS



(B2) All individuals have identical indifference maps. Thus their utility functions are identical up to monotonic transformations. We assume that the utility functions are continuous, strictly increasing in c and δx, and differentiable in c and x. Furthermore, $U(c, 0) > 0$ for $c > 0$, i.e. the telephone good is inessential, also $U(0, x) = 0$ for all $x \geqslant 0$, i.e. the consumption good is essential. It follows that indifference curves do not intersect the x-axis. However, any point on the c-axis is intersected by some indifference curve, but there may be indifference curves not intersecting the c-axis. The former curves will be denoted by $c(x;y)$, where y is the intersection point of the curve with the c-axis, so $y = c(0;y)$ as is shown in Figure 1a, assuming the price of c to be unity. Thus,

$$c(x;y) = \{c\colon U(c, x) - U(y, 0) = 0\}.$$

Finally, we assume that

$$\frac{\partial U}{\partial c}[c(0;y), 0] \quad - \frac{\partial U}{\partial c}[c(x;y), x] < 0,\ y \geqslant 0,\ x > 0. \tag{3}$$

This condition implies that whenever telephone services are consumed, the increment in consumption goods needed to get to a higher indifference curve is smaller than when no such services are consumed.

(B3) Each individual maximizes his utility subject to a budget constraint, thus

$$\max_{c(t+1),\delta} U[c(t+1), \delta x(t)] \quad \text{subject to } c(t+1) + \delta p_x = y(t+1). \tag{4}$$

This specifies the dynamics of the problem: all individuals make their consumption decisions at the beginning of each period on the basis of the telephone stock available at that time, which is the stock of the preceding period. This assumption will be discussed in more detail in Section 4.

We shall now show that given (B2) and in particular (3), individuals will enter the telephone system in the order of their income. Conversely, assume with AA that individuals are utility maximizers and always enter in the order of their income. Assume in addition to AA that utility functions are differentiable (the identity of utility functions for all individuals being implied in their proof), then they must also satisfy (3).

To show the first claim recall the definition of $w(y;p_x)$ as given in (1). Now using (3) and implicit differentiation of (1), it is seen that $w(y;p_x)$ is identical for all individuals and is strictly decreasing in y. Hence, given a telephone stock $\hat{x}$, if $\hat{y} = \{y: \hat{x} = w(y;p_x)\}$, then individuals with income above $\hat{y}$ will join the system and individuals with income below $\hat{y}$ will not. Conversely assume individuals are utility maximizers and enter in the order of their income, but assume $w(y;p_x)$ increasing over some interval. Then it is possible to construct a contradiction by considering a telephone stock x intersecting $w(y;p_x)$ in such an interval. Then, individuals with income just below y prefer a telephone, while those with income just above y do not.[7] Following their preferences, they join the system in the wrong order.

Hence, if at any time t there are $x(t)$ telephones in the system, then let

$$z(t) = \{y: x(t) = N(y), \underline{y} \leqslant y \leqslant \bar{y}\}. \tag{5}$$

Then all individuals above $z(t)$ subscribe to a telephone, but nobody with smaller income does.

The following further assertions can be made regarding the shape of $w(y)$.

Lemma: Assume (B1) and (B2) hold. Then there exists a value $\tilde{y}$ as a function of p_x such that $\tilde{y}(p_x) \geqslant p_x$, $\tilde{y}(0) = 0$ and $\dfrac{d\tilde{y}}{dp_x} > 0$. Furthermore, $w(y;p_x)$ is defined for $y > \tilde{y}(p_x)$, and $w(y;p_x)$ is infinite for $0 \leqslant y \leqslant \tilde{y}(p_x)$ so that an individual in this income bracket will never subscribe to a telephone no matter how large x. For given p_x, $0 < p_x < \infty$, the function $w(y;p_x)$ is continuous and strictly decreasing with $0 \leqslant w(y;p_x) < \infty$, for $\tilde{y}(p_x) < y \leqslant \infty$. Also, approaching $\tilde{y}(p_x)$ from above, $\lim\limits_{y \to \tilde{y}(p_x)} w(y;p_x) = \infty$ and finally, the larger p_x, the higher the curve $w(y;p_x)$, i.e. $\dfrac{\partial w}{\partial p_x}(y;p_x) > 0$ and $w(y;0) = 0$ for all y.

Proof: Above it has been shown that $w(y;p_x)$ is strictly decreasing in y. However, for given y and p_x there may not exist any finite x such that (1) is satisfied. Since $c(x;y)$ is strictly decreasing but does not intersect the horizontal axis by (B2), this can be the case only if $c(0;y) - c(\infty;y) \leqslant p_x$. Now assume $w(y;p_x)$ exists for $y = \hat{y}$, then it exists for $y > \hat{y}$, since by definition of $c(x;y)$ and from (3)

$$\frac{d}{dy}[c(0;y) - c(x;y)] = \frac{\dfrac{\partial U}{\partial c}[c(x;y), x] - \dfrac{\partial U}{\partial c}[c(0;y), 0]}{\dfrac{\partial U}{\partial c}[c(x;y), x]} > 0, x > 0.$$

In particular then, the difference $c(0;y) - c(\infty;y)$ must be strictly increasing in y. But also $c(0;0) - c(\infty;0) = 0$, so there must exist a unique $\tilde{y}$ given by

---

[7] If $w(y;p_x)$ is horizontal over some interval, then no contradiction necessarily exists. But it is still true that in this case an individual with a higher income may join the system though an individual with a lower income does not.

$$\tilde{y}(p_x) = \{y: c(0;y) - c(\infty;y) = p_x\}$$

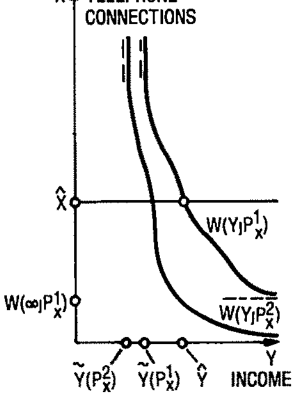
satisfying $\tilde{y}(0) = 0$ and $\dfrac{d\tilde{y}}{dp_x} > 0$. For $y > \tilde{y}(p_x)$, $w(y;p_x)$ must be finite for finite y by its construction and $w(y;p_x) \xrightarrow[p_x \to 0]{} 0$ for all y. Differentiating $w(y;p_x)$ implicitly with respect to p_x, it follows that $\dfrac{\partial w}{\partial p_x} > 0$. By definition of w and $\tilde{y}(p_x)$, $w[\tilde{y}(p_x);p_x] = \infty$. This completes the proof.

All of this may be clarified by some simple graphs. Figure 1a shows a map of indifference curves as explained previously. Each point of the vertical c-axis is intersected by some indifference curve, but the horizontal x-axis is never intersected. Indifference curves, of course, do not need to be asymptotic to the horizontal axis but can be asymptotic to other horizontal lines, depending on the utility level of the indifference curve. Consider an individual with income y. He may spend all income c, thus consuming $(c, x) = (y_1, 0)$ and reaching indifference curve $c(x;y_1)$. Instead $y_1 - p_x$ dollars may be spent on c (assuming $y_1 > p_x$) and p_x dollars on the telephone. This alternative consumption plan is given by some point on the horizontal line $y_1 - p_x$ and will be chosen if $x > w(y_1)$ and rejected if $x < w(y_1)$. The particular x which leaves the individual on indifference curve $c(x;y_1)$ defines $w(y_1)$. By proceeding in the same manner with all y a curve (c, w) is traced.

Also note that if $y < p_x$, then there clearly is no way the line $y - p_x$ could intersect an indifference curve and w is not defined. This may even occur if $y > p_x$ as is shown in Figure 1a. For example, if $y_1 - p_x < c(\infty;y_1)$, then w is not defined and is set equal to infinity. In this case, if the price of a telephone subscription were just equal to $y_1 - c(\infty;y_1)$, then $y_1 = \tilde{y}(p_x)$. Figure 1b shows $w(y)$ as a decreasing function of y for two values of p_x. Given some arbitrary stock of telephones $\hat{x}$ shown as a horizontal line, we know that if $w(y) \leqslant \hat{x}$, individuals will subscribe to a telephone, while none will do so if $w(y) > \hat{x}$. Now, since $w(y)$ is decreasing, we are sure that individuals with higher income will subscribe to a telephone before individuals with lower income do. But the reverse would hold if $w(y)$ were increasing over some interval, hence the importance of condition (3) given under (B2).

FIGURE 1b

NUMBER OF TELEPHONE CONNECTIONS AT WHICH CONSUMERS WITH INCOME Y ARE INDIFFERENT BETWEEN SUBSCRIBING AND NOT SUBSCRIBING



Next, a proposition is proved which corresponds to that given by AA. It is understood in stating and proving it that $N^{-1}(h) = \underline{y}$, $N^{-1}(0) = \bar{y}$, and $w^{-1}(x) = \infty$ if $x < w(\infty)$.

Proposition 1: Let $x(0) = x_0$ be the initial stock of telephones subscribed to at $t = 0$ by individuals with income above $z(0) = N^{-1}(x_0)$, with $h \geqslant x_0 \geqslant 0$, and let (B1)–(B3) be satisfied. Furthermore, assume that $w(y)$ and $N(y)$ are not identical functions over some nonzero interval. Then for future time periods the stock of telephones is given by a unique sequence $\{x(t)\}$, $t = 1, 2, 3, \ldots$ which depends of x_0, such that

$$x(t+1) \left\{ \begin{matrix} > \\ = \\ \leqslant \end{matrix} \right\} x(t) \qquad \text{for all } t, \qquad t = 0, 1, 2, \ldots$$
$$\text{if} \quad w^{-1}(x_0) \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} N^{-1}(x_0). \tag{6}$$

The proof follows from the construction of the sequence. Assume $0 \leqslant x(t) \leqslant h$ and

$$x(t) = N[z(t)], \tag{7}$$

which is true for $x(0)$. Then from (A3) it follows that at $t+1$ individuals with income above $w^{-1}[x(t)]$ decide to subscribe to a telephone, so

$$z(t+1) = \begin{cases} \bar{y} & , \; w^{-1}[x(t)] > \bar{y} \\ w^{-1}[x(t)] & , \; w^{-1}[x(t)] \leqslant \bar{y}. \end{cases} \tag{8}$$

If $x(t) < w(\bar{y})$, then even individuals with income greater than $\bar{y}$ would not want a telephone. Since such individuals do not exist, $z(t+1)$ is set equal to $\bar{y}$, implying that no individual in the population desires a telephone at time $t+1$. Note also that $z(t+1) \geqslant \underline{y}$ since $w^{-1}[x(t)] > \tilde{y}(p_x) \geqslant p_x \geqslant \underline{y}$ by the Lemma and (B1), so $\underline{y} \leqslant z(t+1) \leqslant \bar{y}$. Individuals with income greater than $z(t+1)$ want a telephone, so

$$x(t+1) = N[z(t+1)]. \tag{9}$$

So at $t+1$ a unique telephone stock is determined with $0 \leqslant x(t+1) \leqslant h$. Hence, returning to (7) the sequence is constructed. It remains to show (6). Assume that one of the inequalities or the equality in (6) holds, so using (7) and (8)

$$z(t+1) = w^{-1}[x(t)] \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} N^{-1}[x(t)] = z(t). \tag{10}$$

Both $z(t+1)$ and $z(t)$ are elements of $[\underline{y}, \bar{y}]$ for which $N(y)$ is strictly decreasing, so using (7) and (9)

$$x(t+1) = N[z(t+1)] \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} N[z(t)] = x(t).$$

Here two cases have to be considered. Either both $x(t)$ and $x(t+1)$ are larger than $w(\infty)$ and then it follows that

$$w^{-1}[x(t+1)] \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} w^{-1}[x(t)],$$

since $w^{-1}(x)$ is strictly decreasing. Or one or both $x(t)$ and $x(t+1)$ are smaller than $w(\infty)$. In this case the possibility that $x(t) < x(t+1)$ can be excluded, since then by assumption $x(t) < w(\infty)$, implying $x(t+1) = 0$, a contradiction. Hence $x(t) \geqslant x(t+1)$ and $x(t+1) < w(\infty)$ by assumption. Then both $w^{-1}[x(t+1)]$ and $w^{-1}[x(t)]$ may be infinite and we can only conclude in general that if $x(t+1) < x(t)$, then $w^{-1}[x(t+1)]$

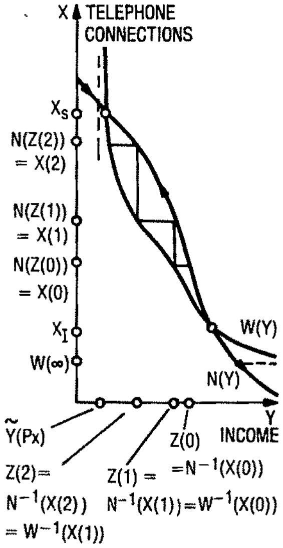
$\geqslant w^{-1}[x(t)]$, a weak inequality. Combining these results with (8) and (9), we have that

$$w^{-1}[x(t+1)] \left\{ \begin{matrix} < \\ = \\ \geqslant \end{matrix} \right\} N^{-1}[x(t)],$$

which proves the proposition.

FIGURE 2

GROWTH AND DECAY OF THE TELEPHONE SYSTEM



The proposition is easily understood from Figure 2, which shows $N(y)$ and $w(y)$ for some p_x. The two curves may or may not intersect and there may be many intersection points. Consider some initial stock $x(0)$ of telephones as shown, then individuals with income at least $w^{-1}[x(0)]$ will subscribe to a telephone in the next period. Since the number of such individuals in the population is known, the stock in period $t=1$ is known, etc. Note that the sequence $\{x(t)\}$ of telephone stocks is dated by period and remains the same independent of the lengths of periods chosen. Hence, it is difficult to associate any real time dimension to this growth process.

Assumption (B1) may be slightly relaxed to allow $\underline{y} > p_x$. In this case Proposition 1 still holds, except that in (6), if $w^{-1}(x_0) < N^{-1}(x_0)$, then $x(t+1) \geqslant x(t)$ for all t, $t = 0, 1, 2 \ldots$, i.e. strict growth cannot be shown anymore.

The proposition may now be compared to that of AA. Our (B1)–(B3) represent a special case of (A1)–(A3). Assumption (B3) is identical to (A4), and in our formulation there is no assumption corresponding to (A5).

Our Proposition 1 characterizes the future behavior of the telephone stock as a function of the initial stock size, rather than an initial change in stock. It is shown that there is either only growth or only decline of the stock of telephones, with stagnation being a possibility in the latter case. Hence, if the exogenous change assumed by AA is identical to the endogenous change implied by their first four assumptions, then no complications arise with the proof of their proposition. However, an arbitrary exogenous increase in telephones may well be followed by a decline in stock size demanded, as is the case in Figure 2, if the upward jump would leave the stock still below x_I, or would carry it beyond x_S. Thus, in order for their proposition to hold true in general, additional assumptions must be made, and in particular qualifications must be given as to the initial stock size assumed. Our proposition more completely describes the possible behavior of telephone demand; in addition, uniqueness is shown.

We furthermore consider the graphical analysis to be useful for an understanding of the nature of this problem. Finally, we would like to emphasize that the functions $N(y)$ and $w(y;p_x)$ are in no way based on the dynamics implied by (B3). So if the dynamics are changed, as we shall do in Section 4, the previous analysis retains much of its usefulness.

3. Comparative statics

■ In this section, we analyze the model introduced above first by considering variations in the subscription price and by specifying the types of equilibria possible and conditions under which multiple equilibria may be obtained. Second, we introduce variations in the size of the population considered and variations in the income distribution. The natural emphasis of the analysis is in deriving conclusions which are due to the public good property of the telephone system. Note that the conclusions will be derived independently of the dynamics implied in (B3).

Let us, for simplicity of exposition, assume throughout that $N(y)$ is differentiable with $y = 0$ and $y < \infty$. Also we continue to assume that $N(y)$ and $w(y)$ are not identical over a nonzero interval. Assumption (B1) with these restrictions added is referred to as (B1.1).

☐ **Variations in subscription price.** The following remarks deal with equilibrium points at which $w(y;p_x) = N(y)$. At any such point total demand for telephone subscriptions, as determined by price p_x and existing subscriptions x, just equals the existing subscriptions. Let (p_x, x) be such an equilibrium point, then

$$\hat{p}_x(x) = \{p_x{:}N(y) = x = w(y, p_x); 0 \leqslant y \leqslant \bar{y}\}. \qquad (11)$$

Observe that (1) $\hat{p}_x(x)$ exists for $0 \leqslant x \leqslant h$ and is unique. To see this note that $N(y)$ is a fixed number for given y, but $w(y;0) = 0$, $w(y;p_x) = \infty$ for some $p_x > 0$ and $\dfrac{\partial w}{\partial p_x} > 0$. Uniqueness follows since $N(y)$ is strictly decreasing on $[0, \bar{y}]$. Furthermore, $N(\bar{y}) = 0$ and $N(0) = h$ imply (2) $\hat{p}_x(0) = 0$ and $\hat{p}_x(h) = 0$, respectively. Also, $\hat{p}_x(x) > 0$ for $0 < x < h$. Since $N(y)$ and $w(y, p_x)$ are continuous functions, it follows that (3) $\hat{p}_x(x)$ is continuous. Finally, (4) $\hat{p}_x(x)$ is bounded from above, for if we let p_x be such that $\tilde{y}(p_x) > \bar{y}$, then $N(y)$ and $w(y, p_x)$ do not intersect. For later reference, let $\bar{p}_x$ be the maximum value of $\hat{p}_x(x)$, $0 \leqslant x \leqslant h$.

It follows from (2) and (3) that $\hat{p}_x(x)$ is an increasing function near the origin. Hence we can state:

Proposition 2: Assume that (B1.1) and (B2) hold. Then there exists an interval $(0, \hat{x})$, $\hat{x} > 0$, such that

$$\hat{p}_x'(x) > 0 \quad \text{for all} \quad x \in (0, \hat{x}). \qquad (12)$$

Such a result would obviously not be possible with a private good. For any normal private good, an expansion of market output under stationary conditions leads to a decrease in equilibrium price. In contrast to that, as a result of Proposition 2, in our case an increase in market output at least in some initial range may be accompanied by a price increase. Thus, if the telephone system is produced by a monopolist, he may be able to expand output and at the same time increase the price of output.

Figures 3a and 3b illustrate the properties of $p_x(x)$. Differentiating $\hat{p}_x(x)$ implicitly by using (11), one obtains

$$\frac{d\hat{p}_x(x)}{dx} \gtrless 0 <=> \frac{dN}{dy} \lessgtr \frac{\partial w}{\partial y}(y;p_x).$$

Using Proposition 1 it follows that if $\hat{p}_x(x)$ is increasing (decreasing), the equilibrium $[\hat{p}_x(x), x]$ is unstable (stable). Also, to any price $p_x < \bar{p}_x$ there must correspond at least two equilibrium telephone stocks.

FIGURE 3a

EQUILIBRIUM STOCK SIZES VARYING WITH TELEPHONE SUBSCRIPTION PRICE

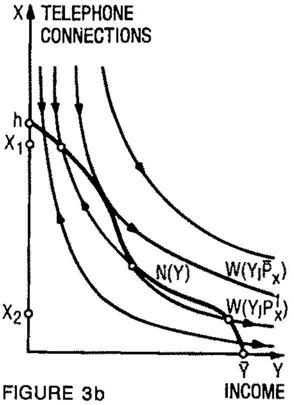
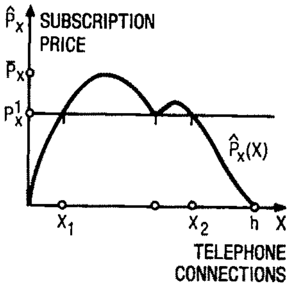


FIGURE 3b

DEMAND FOR TELEPHONE CONNECTIONS



To investigate the properties of equilibrium points further, consider a particular price $p_x < \bar{p}_x$. Then we are sure that $N(y)$ and $w(y;p_x)$ intersect. The number of intersection points must be even, since $N(y)$ and $w(y;p_x)$ are continuous, since all intersection points lie in $[\tilde{y}(p_x), \bar{y}]$, and since $w[\tilde{y}(p_x), p_x] > N[\tilde{y}(p_x)]$ and $w(\bar{y}, p_x) > N(y) = 0$ for $p_x > 0$. Let the set of telephone stocks and incomes corresponding to the set of intersection points be given by $\{x_i\}$ and $\{y_i\}$ respectively, where $i = 1, 2, \ldots, 2n$ with $y_1 > y_2 > \ldots > y_{2n}$ and $x_1 < x_2 < \ldots < x_{2n}$. If there are additional points for which $N(y)$ and $w(y;p_x)$ are tangential without intersecting, these points are not included in the sets.

It is now easy to see that $[\tilde{p}_x(x_i), x_i]$ are alternating stable and unstable equilibrium points with x_1 being unstable. For let y^- (y^+) be an income just below (above) y, then $N(y_1^-) > w(y_1^-, p_x)$ and $N(y_1^+) < w(y_1^+)$, while for y_{2n} the reverse inequalities hold. Similar reasoning shows that y_2 and y_{2n-1}, if they exist, must be stable and unstable respectively, etc.

In addition to $\{x_i\}$ there is a stable point for $x = 0$ and there may be semistable points if $N(y) = w(y;p_x)$ without intersection. A telephone stock of size x_1 represents a minimum threshold size. Disregarding possible semistable sizes below x_1, the telephone stock will fall to zero if it starts from an initial size $x_0 < x_1$. Only if x_0 is larger than the take-off stage x_1 can a positive stock be sustained. On the other hand x_{2n} is a saturation level, which for $y = 0$ and $p_x > 0$ satisfies $x_{2n} < h$. These results are now summarized.

Proposition 3: Assume (B1.1) and (B2) with $0 < p_x < \bar{p}_x$. Then there exists an even number of equilibrium sizes of the telephone system $x_1, \ldots, x_{2n}$, $n \geqslant 1$, ordered by increasing size, half of which are stable and half unstable. There may exist additional semistable points. Stable and unstable points alternate such that if x_i and x_{i+1} are adjacent equilibrium points and x_0 is an initial telephone stock with $x_i < x_0 < x_{i+1}$, then there will be growth if x_i is unstable and decline if x_{i+1} is unstable except possibly if x_0 hits a semistable point. x_1 is unstable and acts as a threshold level. Initial values above this threshold lead to some strictly positive equilibrium stock. Initial values below the threshold lead generally to a zero terminal stock. x_{2n} is stable and acts as a saturation level, $x_{2n} \leqslant h$.

Some immediate policy conclusions may be drawn from this. The existence of multiple equilibria for a given price level implies the possibility of a permanent change in the size of telephone stock without a permanent change in the price level. Hence, given that the telephone stock is at an unstable equilibrium, then it should suffice to lower the price by a small amount for a brief moment to bring the telephone system on a path of self-sustained growth. If the equilibrium is stable instead, and a higher level stable equilibrium exists, then the price level must be lowered for a longer period of time and by a sufficient amount to increase the telephone stock first beyond the intermediate unstable equilibrium. In particular, production will usually have to be subsidized in order to surmount the lower threshold x_1. Finally, observe with respect to the terminal point of convergence x_{2n}, that subsidization of production to lower p_x, or subsidization of the marginal population in order to enable it to enter the telephone system, may increase total welfare because of the welfare gains implied to the rest of the community.

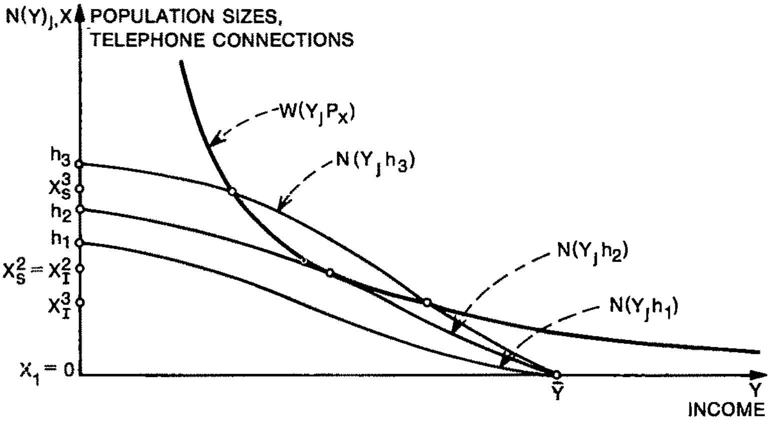
☐ **Variations in population size and income distribution.** In the remainder of this section, we want to trace out the implications of exogenously induced changes of $N(y)$ on threshold and saturation levels. Throughout this analysis we again hold p_x constant. One may motivate this by assuming that the production of telephone connections is at constant cost and the subscription price p_x equals the minimum average cost of producing, maintaining, and servicing a telephone connection. We will mainly analyze the impacts of changes in population at a given income distribution. A thorough analysis of impacts of changes in the income distribution can only be given if $\Psi(y)$ is known. However, some statements can be made about the effects of changes in mean income.

Recalling the definition of $N(y)$ given in equation (1), we may introduce the size of the total population as an argument. Thus, we have

$$N(y;h) = h\,[1 - \Psi(y)], \text{ with } \frac{\partial N}{\delta h} > 0 \text{ for } \underline{y} \leqslant \bar{y} \leqslant y. \quad (13)$$

Consider an initial small population. Suppose, in particular, that the population is of size h_1 with income distribution $N(y;h_1)$ (see Figure 4). Then at given $w(y;p_x)$, there will be no positive equilibrium stock. We can argue that the population h_1, at the given income distribution, is too small to sustain a telephone system. Suppose now that the population increases in number. Eventually a population size h_2 is reached that just sustains the production of a nonzero stock of telephones. Since $N(y)$ approaches $w(y)$ from below, a semistable equilibrium point may be reached, in which there exist exactly x^2 members of the population willing to subscribe to the telephone at a price p_x, given the telephone system already contains x^2 connections. It follows that, in order that this equilibrium point be reached, the system must be established in

FIGURE 4
EQUILIBRIUM STOCK SIZES VARYING WITH POPULATION SIZE



one shot at size x^2. It is obvious from the discussion preceding Proposition 3 that any marginally smaller or larger size of the telephone system will not be sustained.

At any population size beyond h_2 we have, according to Proposition 3, at least two equilibrium points, since for $N(y;h_3)$, $h_3 > h_2$, and given $w(y;p)$, we have $p_x^3 < \bar{p}_x^3$, where $\bar{p}_x^3$ equals the maximum value of $p_x^3(x)$ for given h_3 as defined in the first part of this section. Now it can be shown that, as the total population increases at a given income distribution $\Psi(y)$, not only the total population, but also the proportion of the total population connected to the telephone system increases. Conversely, it will be shown that the simple increase in population at a given income distribution reduces the threshold level for the stock of telephones needed to sustain a growth process; the proportion of the population required to be connected before growth is sustained also reduces. In order to show this, let us redefine the set of equilibrium points x_i by

$$x_i(y;h) = \{x : N(y;h) = x = w(y)\}. \qquad (14)$$

Let, in addition, $n_i = \dfrac{x_i(y;h)}{h}$ define the proportion of the total population h connected to the telephone system, given it is in equilibrium state x_i.

Proposition 4: Assume in addition to (B1.1) and (B2) that $x_i\ (y;h)$ exist. Then for stable equilibrium points x_j, $\dfrac{\partial x_j}{\partial h} > 0$; and $\dfrac{\partial n_j}{\partial h} > 0$; for unstable equilibrium points x_k, $\dfrac{\partial x_k}{\partial h} < 0$, and $\dfrac{\partial n_k}{\partial h} < 0$.

The proof will be established for stable equilibrium points. It is analogous for unstable ones. Substituting (13) into (14) and solving for y, furthermore recalling the discussion preceding Proposition 3 about unstable and stable equilibrium points, we get

$$x_j(y;h) = \left\{x : \Psi^{-1}\left(1 - \frac{x}{h}\right) = w^{-1}(x), \frac{\partial N^{-1}}{\partial x} < \frac{\partial w^{-1}}{\partial x}\right\}. \qquad (14a)$$

Differentiating implicitly, letting $v(x,h) = 1 - \dfrac{x}{h}$

$$\frac{\partial x_j}{\partial h} = - \frac{\dfrac{\partial \Psi^{-1}}{\partial v} \cdot \dfrac{x}{h^2}}{\dfrac{\partial N^{-1}}{\partial x} - \dfrac{\partial w^{-1}}{\partial x}} > 0. \qquad (15)$$

In order to show that $\dfrac{\partial n_j}{\partial h} > 0$, using the definition of n_j,

$$\frac{\partial n_j}{\partial h} = \frac{\dfrac{\partial x_j}{\partial h} \cdot h - x_j}{h^2}.$$

Substituting (15), factoring, and cancelling yield the desired result.

Obviously again, this result would not be possible for a private good. In contrast, we would expect for a private good that, as the population increases at a given income distribution, the demand increases absolutely, and in proportion to the population increase.

The impacts of variations in the income distribution, $\Psi(y)$, at a given population size can only be analyzed thoroughly if a particular form of Ψ is assumed which we do not want to do in this paper. It can be verified without difficulty, however, that changes in mean income at given population and given income variance produce effects which are quite similar to the ones stated in Proposition 4 for changes in population size. If equilibrium points exist, then as mean income increases, stable equilibrium points and, in particular, the saturation level increase and instable equilibrium points, in particular, the threshold level, decrease.

4. Alternative dynamics

■ In this section the dynamics assumed in (B3) are discussed and an alternative is developed. It has been shown in Proposition 1 that there exists a unique set of telephone stock adjustments. These adjustments are independent of the period length chosen. Hence, referring back to Figure 2, starting with an initial stock $x(0)$ the telephone stock will always be $x(2)$ after two periods, independent of period length. Indeed, if the period is very small so as to approach zero, the entire adjustment process will be instantaneous; it then takes zero time to do the infinite steps leading from $x(0)$ to x_s. The growth path in real time is therefore extremely dependent on the period length chosen, with infinite growth rates resulting from a continuous analysis.[8] In order to avoid such a result, stock adjustments must be made dependent on period length. One such formulation is developed in this section. Before turning to this, however, it may be useful to seek some justification for the dynamics implied in (B3) and similarly in (A4).

Assumption (B3) implies that there exists some recurrent fixed date, identical for all individuals at which the telephone stock is

---

[8] In the work of AA [3] an attempt has been made to obtain for a specific example a continuous growth path for the telephone stock. In particular, assuming the separable utility function

$$U(c, \delta x) = U_1(c) + \delta[(1 + ah)x - ax^2],\ a > 0, a < \frac{1}{h},$$

they obtain in footnote 11 the expression

$$(1 + ah)x - ax^2 = x + dx.$$

Letting $a = a_1 \Delta t$, as suggested by the reviewer, this expression does indeed lead to the differential equation:

$$\frac{\dot{x}}{x} = a_1(h - x).$$

However, this reinterpretation of a leads, in our opinion, to difficulties with the utility function. For a given level of c and x the utility flow will vary, depending on the arbitrary time period considered. In addition, the utility function given by $U(c, \delta x, \delta \Delta t)$ would, in our view, not represent a special case of the general model.

readjusted. This may be explained by rigidities on either the demand or the supply side. Hence, it may be assumed that all individuals review their consumption decision only at fixed recurrent dates identical for all of them. One explanation for such behavior draws from the public good property of the telephone system and the informational requirements for the individual's consumption decision. Not only must he know the price of the telephone subscription, but information about who already holds a telephone subscription. Hence, one may think of the government or some other source as compiling periodic statistics on the telephone stock in the population. On the date at which updated information is published, for example, in the form of a new telephone book, each individual updates his consumption decision. Between such dates the old decision stands.[9]

Alternatively, there may be rigidities on the supply side. The telephone company may accept applications for or install telephones only at fixed dates. None of these explanations can be considered entirely convincing, however.

In the following, therefore, other dynamics are assumed while (B1) and (B2) and hence functions $N(y)$ and $w(y)$ are retained. It is assumed here that individuals adjust their preferred action at discrete random intervals, but these may differ among individuals. As was suggested before, an individual cannot be assumed to have full information at any time about the number of individuals who hold a telephone subscription. Hence, we distinguish between the telephone stock size $x^p(t)$ perceived by an individual and the actual stock $x(t)$ at time t. We now want to model the fact that information diffusion or an individual's learning process takes time by assuming that $x^p(t)$ is updated at stochastic dates $t_0, t_1, \ldots$. This leads to the following assumption's replacing (B3).

(B3.1) Individuals maximize their utility U at any time on the basis of the telephone stock $x^p(t)$ perceived by them subject to a budget constraint, hence

$$\max_{c(t),\delta} U[c(t), \delta x^p(t)], \quad \text{subject to } c(t) + \delta p_x = y(t). \qquad (16)$$

Here $x^p(t)$ is updated stochastically by individuals independent of each other. If t_i and t_{i+1} are such dates, then $x^p(t) = x(t_i)$ for $t \in [t_i, t_{i+1}]$. Updates occur at a constant rate μ per year and therefore there is an exponential time-to-update distribution.

This assumption does not itself state the dynamic equation but it can easily be derived from here. Note first that an individual will not change his decision unless he has received new information. "Learning," as implied by (B3.1), occurs in jumps. Between jumps there is no change in the state of information. Also note that during each year μh individuals update and possibly revise their decision. These individuals are, of course, randomly distributed among the total population. Hence, we cannot be sure that individuals with high income will enter the telephone system before individuals with low income. There will always be individuals with high income with no updated information who failed to join the telephone system previously. Similarly there may be individuals with low income who should not be telephone subscribers given knowledge of the actual stock.

---

[9] Still, Arrow and Hahn [2], p. 50, point out in a similar context, "In an economy with many agents, however, it seems unnatural to suppose that all do not only adjust their preferred action at discrete intervals, but they do so at the same moment." In the period analysis criticized by them, prices assume the role the telephone stock has in the present case.

Let the telephone stock at time t be $x(t)$. This stock is rented by individuals with income above and below $w^{-1}[x(t)]$, say in the exogenously given proportion η and $(1-\eta)$, respectively, $0 \leqslant \eta \leqslant 1$. To derive a dynamic equation observe that there are two subgroups of individuals which in the event of updated information will change their consumption behavior. First, there are $(1-\eta)x(t)$ individuals with income $y < w^{-1}[x(t)]$ who would leave the telephone system on the basis of new information. Second, there are $N(w^{-1}[x(t)]) - \eta x(t)$ individuals with income $y \geqslant w^{-1}[x(t)]$ who do not yet own a telephone but would want it in case of an update. The two complementary groups which make up the rest of the population h would instead stay with the old decision. Hence, the total change in telephone stock for period $(t, t+dt)$ is given by

$$dx = \mu\{-(1-\eta)x(t) + N(w^{-1}[x(t)]) - \eta x(t)\}\, dt,$$

so that the terms involving η cancel out. Hence the change in telephone stock per year at time t is proportional to the difference between the actual stock and the desired stock given perfect information for all individuals. It follows that the dynamic equation is given by

$$\dot{x}(t) = \mu\{N(w^{-1}[x(t)]) - x(t)\}, \quad x(0) = x_0. \qquad (17)$$

The integration into our previous analysis is easily possible. Referring to Figure 2 we see that previous dynamics imply jumps from x_0 to x_1 to x_2, etc. The change per period is given by the vertical difference between the functions $N(y)$ and $w(y)$ measured at the point where $x(t)$ intersects $w(y)$. Similarly now growth per year is proportional to the same vertical difference. Note, however, that if the current dynamic equation is cast in discrete terms, then a change in the period length from a year to two years must be accompanied by a corresponding change in μ, so the length of the period matters. The next result is immediate and closely resembles Proposition 1.

Proposition 5: Let $x(0) = x_0$ be the initial stock of telephones, with $0 \leqslant x_0 \leqslant h$; let (B1), (B2), and (B3.1) be satisfied; then the growth path of the telephone system is given by (16) and $\dot{x}(t) \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} 0$ for all t if $N[w^{-1}(x_0)] - x_0 \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} 0$.

Since the functions $N(y)$ and $w(y)$ are not affected by the change in dynamics, all the results of Section 3 still hold, in particular, Proposition 3's reference to the points towards which the telephone system may converge.

So far, we did not consider explicitly the role prices may play in the dynamics of the model. The price may control two properties of the telephone stock growth path. First, it limits the size to which the telephone system may finally grow as was discussed previously. Second, it determines how fast the equilibrium is reached. Low initial prices not only decrease the minimum threshold size, but also speed up the growth of the telephone system. Hence, a monopolist or governmental agency operating the telephone system has at any moment the tradeoff between somewhat lower current unit revenues due to a lower price p_x, implying a faster accumulation of stock and reaching a given revenue goal at an earlier time, and a higher price, implying a slow growth process. Assuming that expansion of the existing telephone system is possible at constant unit cost c, the problem for the profit maximizing firm therefore becomes

$$\max \int_0^\infty \{p_x x - (N[w^{-1}(p_x, x)] - x)c\} e^{-\rho t}\, dt,$$

subject to (17) and using p_x as a control variable.

By assuming special forms of $\Psi(y)$ and $U(c, x)$, it is possible to derive properties regarding the price path and the resulting growth path of the stock.

5. Urban growth analogies

■ In the paradigm introduced by Artle and Averous, which we have developed further in the preceding sections of this paper, the telephone system is used as an example to discuss aspects of the demand for a commodity which seems to have a rather peculiar and unique combination of private and public good properties. The cumulative public good property of the commodity in question is, as we recall, due to its characteristics as a stock, and as a private good. Any individual who wants to consume services from this stock must subscribe to get access to it. But due to his subscription, the stock size will be increased and this, in turn, increases all other subscribers' utility.[10]

There are a number of instances, however, in which we can talk about commodities with a combination of properties sufficiently similar to allow a direct reinterpretation of the paradigm given. The example given here refers to the urban economy and its growth in time. As we want to keep the analogies to the above model as close as possible, our example provides only a rather crude description of urban growth. Hence, it should not be taken at face value, but rather as suggestive for further research in this area.

Face to face communication in urban areas has a number of consumption features in common with communication via a telephone system. High population density and the existence of special private and public facilities at a particular location allow for contacts considered important in the conduct of business and pleasure.[11] The mere designation of a specific location as an urban center, however, is not a sufficient incentive for individuals to bear the cost of commuting to the center. Rather, it is the knowledge of the presence of other people there that makes a trip worthwhile. Without consideration of the possibility of overcrowding, the utility for each individual may very well be increasing as the number of other people commuting to the center increases. Alternatively, and perhaps more realistically, it may be argued that it is the existence of private and public services that can be sustained only by the joint consumption of many other people.

---

[10] Thus, "free rider" problems, which seem to be inherent in "Samuelsonian" public goods, are excluded in this setup of the paradigm. However, an individual's consumption of services is most probably not proportional to the number of consumers connected to the system, as assumed above and mentioned in footnote 5. Explicit consideration of this aspect may introduce the problems elucidated upon by Samuelson [8]. If the individual's consumption can be metered, then a two-part tariff is needed, and, in effect, there will be no problem in obtaining efficiency solutions for telephone service provision. If this is not the case, then for the population segment having access to the good, the good is a purely "public" one. In any rate, to take account of such effects, access to and consumption of the service should be represented by separate arguments in the consumer's preference function.

Now, in perfect analogy to the case of telephone consumption, an individual living at some distance from the urban center may derive utility from face to face contacts, and consumption of centrally located services, only if he indeed decides to commute at the expense of a commuting fare.[12] Moreover, his utility increases as more people do commute, or the service level increases. Any individual who indeed decides to commute may then be considered an inhabitant of the urban area.

In this analogy the role of the income distribution in the telephone model is taken up by the distribution of population according to distance from the urban center. Assuming, as before, that all people have identical preferences and, in addition, identical nominal income, and that the transportation cost is strictly increasing with distance, individuals enter the urban system in the order of distance from the center.

As should be apparent by now, it is easy to construct functions $N(y)$ and $w(y)$ with y being the distance from the center, such that the functions can be given the same interpretation as they had in the telephone model. It follows that we can directly infer the following results from the Lemmas and Propositions of the previous sections.

(1) At a given income level and transportation cost function, there exists a distance $\tilde{d}$ such that any individual living at or beyond that distance is not willing to commute, no matter how large the city grows in terms of number of commuters. Clearly, that distance becomes smaller as the transportation cost increases. Also, as transportation cost increases, an individual is willing to commute only if the number of other commuters increases.

(2) If individuals living near the center have a nonzero commuting cost, then there exists an initial "threshold of commuters," or a threshold city size which must be overcome to set the growth process into motion. Any initial size below this minimum size threshold will not be sustained.

---

[11] See Hoover and Vernon [6], in particular, Chapter 3, pp. 58–73, Artle [3], pp. 65–78, and Alonso [1], pp. 23–24.

[12] In models of urban location, it is typically assumed that people move to the city center only for work. The explanation given in this model may be considered complementary. It may eventually hold better for cities outside the United States, however.

(3) There is a definite terminal city size, given income and transportation cost. There might, however, be intermediate stable city sizes, so several thresholds may have to be surmounted to bring the system up to the terminal state. Such intermediate stable sizes may be caused either by humps in the transportation cost function, as for example produced by rail based mass transit systems, or by the fact that there are less densely populated intermediate distance intervals.

(4) As the total population in the region considered increases at any distance in proportion to the actual population density, the minimum size threshold required to induce growth is reduced, and the terminal city size in terms of radius and commuting population is increased.[13]

It remains to take explicit account of the dynamics involved in the growth process. For this model, the most appropriate argument may be that growth is constrained on the supply side. It may be argued that time is needed to establish the transportation infrastructure so that the commuting radius can be increased.

In conclusion, we would like to mention that a similar analogy can be made for the growth of productive capacity in urban areas. It is commonly argued that there are economies of scale related to the existence of a large bulk of production at one locality that are external to any one individual firm taken in isolation.[14] As the productive capacity in the city increases, the economies external to the individual firm increase, thus raising the marginal productivity of labor and urban wage rate. Hence, laborers located at some distance from the city are induced to commute to the urban area if the wage net of commuting costs exceeds the wage they can earn at their place of living. Again, if an initial size threshold is surmounted,[15] a self-reinforcing growth process will start analogous to the one analyzed for the telephone system.

References

1. Alonso, W. "Industrial Location and Regional Policy in Economic Development." Working Paper No. 74, University of California, Berkeley: Institute of Urban and Regional Development, Center for Planning and Development Research, February 1968.
2. Arrow, K. and Hahn, F. H. *General Competitive Analysis.* San Francisco: Holden-Day, Inc., 1971.
3. Artle, R. "Urbanization and Economic Growth in Venezuela." *Regional Science Association: Papers and Proceedings,* Vol. 27, 1971.
4. ——— and Averous, C. "The Telephone System as a Public Good: Static and Dynamic Aspects." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 89–100.

---

[13] In this simple model, Proposition 2 has no direct analogy, for we have not introduced any fixed cost regulating access to the urban area.

[14] See for example Hoover [5], pp. 78–81, and especially Chapter 8, and Alonso [1].

[15] This requirement, derived from the analogy to the telephone model, seems to confirm observations of development economists such as Rosenstein-Rodan [7]. He recommends providing a once-and-for-all subsidy which, if applied in a "balanced" form to the industries concerned leads to a take-off for urban, or regional growth.

5. HOOVER, E. *The Location of Economic Activity*. New York: McGraw-Hill, 1948.
6. ——— and VERNON, R. *Anatomy of a Metropolis.* New York: Anchor Books, 1962.
7. ROSENSTEIN-RODAN, P. N. "The Theory of the Big Push," in H. Ellis, ed., *Economic Development of Latin America,* New York: The Macmillan Company, 1961, p. 59 ff.
8. SAMUELSON, P. A. "Pure Theory of Public Expenditure and Taxation," in J. Margolis and H. Guitton, eds., *Public Economics*, London: Macmillan, 1969, pp. 98–123.

Insurance as a market in contingent claims: structure and performance

John M. Marshall

Assistant Professor of Economics
University of California, Santa Barbara

Insurance considered as brokerage is very different from insurance considered as holding funds. The scale economies which are ubiquitous in the funds model dissipate when insurance is considered as intermediation. There appear to be no barriers to competition, for the potential entrant has no need to achieve large size, although economies of internal clearing give firms an incentive to merge. The alternative to insurance in the brokerage model is a combination of incorporation and reinsurance. Even in the presence of a state-imposed cartel this alternative limits the possible scope of insurance price manipulation. Contrary to the view of the funds theory—that insurance profitability depends on holding and safeguarding an appropriate fund—here capital market prices are the determinants. Since capital market prices are not in the control of insurance regulators, their intervention in setting insurance prices leads to gaps in insurance coverage.

1. Introduction

■ The most widely accepted theory of insurance was formulated by Alan Willett, who in 1901 defined insurance as the accumulation of reserves against contingencies. He writes that "the term insurance . . . is . . . used in describing the fund accumulated to meet uncertain losses."[1] The insurance industry exists because of economies of scale inherent in reserve accumulation.[2] While little is written about the reserves concept as economic theory, it is widely applied in the insurance literature, in both textbooks and periodicals. The concept is well suited to insurance education, for it centers on the operation of the firm and the management of insurance funds.

Another current theory of insurance describes it as trade in contingent claims.[3] This approach passes over the fund and the management of insurance firms while stressing the reciprocity and mutuality of insurance contracts. Much has been written about this conception as pure theory, often with attention to financial markets, but contributions applying and interpreting the theory for the insurance industry are few. The reason for the lack of applications

---

The author wishes to thank Perry Shapiro for numerous valuable suggestions.

[1] In [6], p. 71.

[2] *Ibid.*, Ch. 7.

[3] The main statements of the theory are familiar from Arrow [1] and from many others. It has been rephrased for insurance by Marshall [5], Kihlstrom and Pauly [4], and Brainard and Dolbear [2].

to the industry is clear; in the basic contingent claims model the industry has no role and no reason for being. A contingent claims model with an endogenous insurance industry is needed. It should accommodate the analysis of insurance regulation as well as describe the relationship of insurance to other trade in contingent claims. The model offered here does that, at least in a limited fashion, by viewing transactions costs as the production costs of intermediaries, among which insurance firms have a special role.

2. The contingent claims model without transactions costs

■ Suppose that there are individuals $j = 1, 2, \ldots, n$, and possible future states of nature $s = 1, 2, \ldots, S$. For each state individual endowment is $\overline{Y}_s^j$. Individuals hold subjective probabilities π_s^j for states s. Before it is known which state will occur, there is trade in claims contingent upon the outcome. Individual consumption, contingent upon the occurrence of s, is Y_s^j. Individual preferences are expressed in the index $V^j = \sum_s \pi_s^j U_s^j(Y_s^j)$, where U_s^j is an index of preferences conditional upon the occurrence of state s. The index V^j is quasi-concave in the Y_s^j. This would be a consequence of having each U_s quasi-concave but would not necessarily imply it.

Markets are perfect in the usual senses: no individual is conscious of being able to affect prices, and there are no transactions costs. Given prices $P_1, P_2, \ldots, P_S$, each consumer must

$$\text{Maximize } \sum_s \pi_s^j U_s^j(Y_s^j) \tag{1}$$

$$\text{subject to } \sum_s P_s(Y_s^j - \overline{Y}_s^j) = 0. \tag{2}$$

The market clearing conditions are

$$\sum_j Y_s^j = \sum_j \overline{Y}_s^j \qquad \text{for all } s. \tag{3}$$

Under the assumption that no individual is satiated at any attainable consumption and with the observation that the endowments are fixed and finite, an equilibrium price system exists[4] and the equilibrium is a Pareto optimum.[5] A Pareto optimum means that no rearrangement of contingent claims could raise any consumer's *ex ante* satisfaction without lowering another consumer's *ex ante* satisfaction. The result is critically dependent upon the lack of market imperfections, and therefore it rules out the use of resources to make exchanges.

3. Brokerage as production

■ **Transactions costs.** Transactions costs are introduced in the basic model as the production costs of brokerage firms. Brokerage firms arise naturally in this pure exchange model with transactions costs because of the various economies associated with volume in making transactions. Volume economies include indivisibility; a certain fixed input is needed to make any transaction, however small. There are also economies realized in grouping transactions in advantageous ways; the total cost of making many associated transactions together may be less than the cost of making them separately. Finally, there may be residual economies of scale which lower the average cost of transaction after the influence of the initial indivisibilities is past. For these reasons it is uneconomical for each individual to carry on his own transactions. Firms arise that specialize in intermediation and offer their services at costs less than those individuals would face if they made the transactions themselves.

---

[4] By application of Theorem (1) of Debreu [3], chapter 5.7.

[5] Apply Theorem (2) of Debreu [3], chapter 6.3.

These observations justify the existence of brokerage firms, but it is not clear whether they dictate the cost structure of firms operating in a relevant range. It is possible, although not necessary, to suppose that these volume economies and indivisibilities are exhausted at small scale, and that, thereafter, the marginal cost and average cost of transactions are the same. In describing intermediation it is useful to assume that the basic transaction is made with constant cost. It follows in the analysis that various economies are still available to firms, especially insurance firms, which operate in several basic transactions at once.

There are three separate markets for contingent claims. The market in which consumers sell and brokerage firms buy is called the W-market, with prices P_s^w; the market in which brokerage firms sell and consumers buy is the Z-market, with prices P_s^z. Since firms only buy in the W-market and only sell in the Z-market, $P_s^w \leqslant P_s^z$. The size of the price differential depends upon the technology of brokerage and upon the organization of the brokerage industry.

Brokerage firms are not required to bear risks. They arrange mutual risk-sharing among consumers by balancing W-market purchases with Z-market sales. To facilitate this balancing, brokerage firms have a reinsurance market in which they trade contingent claims among themselves. Ordinary consumers are excluded from the reinsurance market by prohibitive transactions costs or institutional constraints. Like the W-market and Z-market, the reinsurance market has no transactions costs. There are, however, transactions costs involved when purchases in the W-market are matched with reinsurance market sales or Z-market sales are matched with reinsurance market purchases. This is completely analogous to the treatment of transactions costs in the matching of W-market purchases with Z-market sales.

The reinsurance market is part of the system of financial markets. This analysis abstracts from all differences and separations within that system. The reinsurance market is hereafter called the "capital market." Prices in the capital market are $P_1, P_2, \ldots,$ P_S.

An illustration of the principles involved may be needed at this point. Consider two consumers, One and Two. Each has a single loss which may or may not occur. Let state one ($s = 1$) be the state in which neither loss occurs, state two ($s = 2$) be the state in which only Two's loss occurs, state three ($s = 3$) be the state in which only One's loss occurs, and state four ($s = 4$) be the state in which both losses occur. Contingent claims are created by writing contracts for states as follows:

If $s = 1$, IOU one bushel of goods (Individual One).
If $s = 3$, IOU one bushel of goods (Individual One).
If $s = 1$, IOU one bushel of goods (Individual Two).
If $s = 2$, IOU one bushel of goods (Individual Two).

These represent states in which One and Two wish to offer contingent goods.

Taking these contracts to market, One and Two sell to intermediaries at the W-market. The intermediaries carry the contracts to the Z-market, organize them in portfolios that One and Two desire to buy, and resell the contracts there. After this transaction One has some of Two's obligations for state 3 and Two has some of One's obligations for state 2. In making up the portfolios of contingent claims that One and Two want, the brokerage takes state one goods to the capital market where it sells them for state four goods. Then both One and Two purchase contingent claims in state four, as they desire to do.

☐ **Intermediaries as producers.** The brokerage firm buys quantities W_s in the W-market and sells quantities Z_s in the Z-market. Under certain conditions it may also resort to trade in the capital market where its purchases are X_s^- and its sales are X_s^+. The quantity w_s is transferred from the W-market to the Z-market, where w_s is the excess of W_s over X_s^+. All of these quantities are constrained to be nonnegative.

Transactions costs associated with transfer from the W-market to the Z-market are $A(\sum_s P_s w_s)$; those associated with transfer from the W-market to the capital market are $B(\sum_s P_s X_s^+)$; and those associated with transfer from the capital market to the Z-market are $C(\sum_s P_s X_s^-)$. Costs must be paid in each state, and there are no scale economies for basic transactions. Thus $A = a \sum_s P_s w_s$; $B = b \sum_s P_s X_s$; and $C = c \sum_s P_s X_s^-$. It is further assumed that in the initial equilibrium of the capital market all prices P_s are equal. This can be achieved artificially by supposing that the prices, although unequal, are rational numbers, $P_s = m_s/n_s$, where m_s and n_s are integers. Let n^* be the least common multiple of the n_s and rewrite $P_s = m_s'/n^*$. Conceptually each state s may be divided into m_s' equal parts, each with price $1/n^*$. In spite of the possibility of this transformation, the assumption of equal prices puts a real restriction on the structure of transactions costs. If equal prices are assumed, $A = a \sum_s w_s$, $B = b \sum_s X_s^+$, and $C = c \sum_s X_s^-$, where the constant value P_s has been absorbed by the multiplicative constant. It is required that $1 - Sa > 0$, $1 - Sb > 0$, and $1 - Sc > 0$.

Vector notation is employed in the derivations that follow. Vectors are distinguished by lack of subscripts as $W = (W_1, W_2, \ldots, W_S)$, $Z = (Z_1, Z_2, \ldots, Z_S)$, etc. Let $\Gamma(x)$ denote the $S \times S$ matrix with x at every entry, and let $\Phi(x)$ be $\Gamma(x)$ plus the identity matrix. It is readily verified that $\Phi^{-1}(x)$ is equal to $\Phi[-x/(1 + Sx)]$ and that $\Phi(x)\Phi(y)$ is equal to $\Phi(x + y + Sxy)$. In this notation the cost of transferring a vector w of contingent goods from the W-market to the Z-market is $\Gamma(a)w$, and the net quantity available in the Z-market as a result of the transfer is $\Phi(-a)w$. The symbol $x \geqslant y$ for vectors x and y indicates each entry of $x \geqslant$ the corresponding entry of y.

With this notation the brokerage firms' production constraints are

$$W \geqslant X^{+} + w \tag{4}$$

$$\Phi(-a)w + \Phi(-c)X^{-} \geqslant Z + \Gamma(b)X^{+}. \tag{5}$$

Equation (4) states that the total purchases of contingent goods in each state, W_s, must at least equal the quantity of contingent claims for that state sold in the capital market, X_s^{+}, plus the quantity of contingent goods transferred to the Z-market, w_s, for use there. Similarly, equation (5) states that the sources of contingent goods available at the Z-market must be at least as great as the uses of contingent goods. Resources are $\Phi(-a)w$, which is the vector of contingent goods arriving in the Z-market when w is the vector sent from the W-market, and $\Phi(-c)X^{-}$ is the vector of contingent goods arriving at the Z-market if X^{-} is the vector of contingent claims sent from the capital market. In both these quantities transactions costs are already netted out. The uses of contingent claims are the vector Z of sales and the vector $\Gamma(b)X^{+}$, which represents transactions costs paid to transfer the vector X^{+} of contingent claims from the W-market to the capital market. In formulating a model such as this one it is necessary to specify where transactions costs are borne. Here resources must be purchased and transferred to the Z-market before they can be disbursed to pay transactions costs.

In general, a brokerage firm will wish to maximize profits, that is

$$\text{Maximize} \quad P \cdot (X^{+} - X^{-}) + P^{z} \cdot Z - P^{w} \cdot W, \tag{6}$$

subject to the constraints of equations (4) and (5), and

$$X^{+} \geqslant 0, X^{-} \geqslant 0, w \geqslant 0, Z \geqslant 0, W \geqslant 0. \tag{7}$$

This completes the definition of the general brokerage form.

☐ **Consumers.** Denote a consumer's W-market sales as W_s^j, and his Z-market purchases as Z_s^j. Then the consumer maximizes utility subject to a budget constraint, that is

$$\text{Maximize} \quad \Sigma\, \pi_s^j U_s^j(\overline{Y}_s^j + Z_s^j - W_s^j) \tag{8}$$

$$\text{subject to} \quad P^{w} \cdot W^{j} - P^{z} \cdot Z^{j} \geqslant 0. \tag{9}$$

This differs from the consumer's problem in the zero-transactions-costs world only in the form of the budget constraint. The budget constraint is no longer a plane but a polyhedron with vertex at the consumer's endowment point. The rate of exchange between two commodities depends upon which one the consumer is buying and which one he is selling.

☐ **Comparison to financial markets.** Trade in financial assets other than insurance could be introduced in the same way. Qualitatively the analysis is identical, but there is reason to believe that the level of transactions costs is much lower for ordinary financial assets. This is seen by considering the degree of diversity and specificity of insurance contracts as opposed to financial assets. The states, s, are defined over a probability space which will be called E. A partition of E is a collection of mutually disjoint subsets of E which collectively exhaust E. For insurance transactions, the probability space is partitioned by $\underline{S} = \{A_1, A_2, \ldots, A_S\}$, where A_s is now the outcome previously indexed simply by s. Consider another partition $\underline{R} = \{B_1, B_2, \ldots, B_R\}$ of E. Partition $\underline{R}$ is strictly coarser than partition $\underline{S}$; that is, if B_r is an element of $\underline{R}$, then $B_r = \cup A_{si}$, $A_{si} \in \underline{S}$, and $\underline{R} \neq \underline{S}$. $\underline{R}$ is the set of outcomes on which financial asset transactions can be made. Transacting contingent claims for outcomes specified by $\underline{R}$ is cheaper than transacting contingent claims for outcomes in $\underline{S}$.

A separate group of brokerage firms carries on trade in financial assets. These firms use the capital market as their wholesale market. Because transactions costs are lower in financial assets, the prices faced by consumers are nearer those of the capital market. The price advantage is offset, however, by the inferior diversity and specificity of financial assets. Consumers trading in both markets at once use the financial market to diversify their investment portfolio and use the insurance market to insure their personal risks. In ignoring trade in financial assets, the formal model of this paper bypasses an important aspect of consumer behavior under risk. This aspect of consumer behavior and the relation of the insurance system to other financial markets are significant areas needing more research.

4. Insurance as a form of brokerage

■ A brokerage is technically able to perform transactions among the three markets. A brokerage is defined to be an insurance brokerage if it must satisfy a barter constraint

$$P^z \cdot Z - P^w \cdot W = 0. \tag{10}$$

This constraint restricts the insurer to trading contingent claims among consumers and has a basis in the licensing and regulatory requirements for insurance companies. Two other kinds of specialized brokerage exist, which perform functions forbidden to insurance companies. An incorporator is a firm which specializes in transactions between the W-market and the capital market. An incorporator sells consumers contingent goods in the capital market, but is constrained against selling them back to other consumers; that is, $Z = 0$. A reinsurer specializes in transactions between the capital market and the Z-market. A reinsurer buys contingent claims in the capital market on behalf of customers in the Z-market but is constrained against purchasing contingent claims from other consumers; that is, $W = 0$. The constraint on insurers implies, from equation (6), that the insurer's objective is to maximize $P \cdot (X^+ - X^-)$, which is the capital market valuation of the insurer's capital market trade. In a world without transactions costs this would be equivalent to maximizing the present discounted value of shareholders' wealth and would be consistent with the optimization of their consumption. Transactions costs threaten to alter this criterion. Shareholders may wish to trade their portions of contingent profit and they incur transactions costs in doing so. Optimizing the shareholder's welfare requires giving him a package of contingent claims which augments his portfolio without forcing him to make costly trades. Since capital market transactions are costless to brokerages, it is assumed that they provide each individual shareholder with precisely the contingent claims he wants. Then the value maximization criteria is again consistent with optimizing shareholder's welfare.[6]

☐ **Economies of internal clearing.** Transacting contingent claims between the W-market and the Z-market can be done in two ways. A vector W of contingent claims can be transferred directly to the Z-market with proceeds $\Phi(-a)W$. The same vector of contingent claims can be transacted through the capital market, where it yields proceeds $X^{+} = W = X^{-}$. The proceeds of this transaction in the Z-market are $\Phi(-c)X^{-} - \Gamma(b)X^{+} = \Phi(-c-b)W$. When $a < b + c$, the direct transaction is cheaper than the one using the capital market. In this situation economies of internal clearing are said to exist. The case of $b + c < a$, that is, external clearing cheaper than internal clearing, is uninteresting. In this case all transactions between the W- and Z-markets pass through the capital market, and it may be assumed that $a = b + c$. Thus, $a \leqslant b + c$.

The existence of economies of internal clearing has important implications for the operation of general brokerage firms. Brokerage firms act to solve the problem described in equations (4) through (7). Associated with the solution to this problem are nonnegative constants λ_s and δ_s. The former is the marginal value of W_s to the firm, the latter is the marginal cost of providing Z_s. These constants are interpreted as the prices in the W-market and Z-market at which the firm can just afford to demand or supply contingent goods. The constraints in equation 7 imply the existence of multipliers Ψ_s, θ_s, ϕ_s, η_s, and ϵ_s, where $\Psi_s = 0$ if $X_s^{+} > 0$, $\theta_s = 0$ if $X_s^{-} > 0$, $\phi_s = 0$ if $w_s > 0$, $\eta_s = 0$ if $W_s > 0$, and $\epsilon_s = 0$ if $Z_s > 0$, and nonzero values of the multipliers represent the marginal costs of the corresponding constraints when they are binding.

For clarity, the Lagrangian format is used to derive Kuhn-Tucker conditions for optimization. This is equivalent to use of the dual of the linear program, but seems more natural here. Then it is required to find a saddle point of

$$\begin{aligned} L = P \cdot (X^{+} - X^{-}) + P^{z} \cdot Z - P^{w} \cdot W + \lambda \cdot (W - X^{+} - w) \\ + \delta \cdot [\Phi(-a)w + \Phi(-c)X^{-} - Z - \Gamma(b)X^{+}] \qquad (11) \\ + \Psi \cdot X^{+} + \theta \cdot X^{-} + \phi w + \eta \cdot W + \epsilon \cdot Z. \end{aligned}$$

First-order conditions, in addition to the constraints 4, 5, and 7 are

---

[6] It would be simpler but less satisfying to assume that brokerages are constrained by institutional structure to sell their profits in the capital market and transfer the monetary proceeds to shareholders. This, too, would restore the value maximization criterion.

$$P - \lambda - \Gamma(b)\delta \leqslant -\Psi, \quad \text{where } \Psi \cdot X^+ = 0 \text{ and } \Psi \geqslant 0 \tag{12}$$

$$-P + \Phi(-c)\delta \leqslant -\theta, \quad \text{where } \theta \cdot X^- = 0 \text{ and } \theta \geqslant 0 \tag{13}$$

$$-\lambda + \Phi(-a)\delta \leqslant -\phi, \quad \text{where } \phi \cdot w = 0 \text{ and } \phi \geqslant 0 \tag{14}$$

$$-P^w + \lambda \leqslant -\eta, \quad \text{where } \eta \cdot W = 0 \text{ and } \eta \geqslant 0 \tag{15}$$

$$P^z - \delta \leqslant -\epsilon, \quad \text{where } \epsilon \cdot Z = 0 \text{ and } \epsilon \geqslant 0. \tag{16}$$

If internal clearing is strictly cheaper than transactions through the capital market, i.e., if $a < b + c$, then at the optimum *it is impossible for* $X_r^+ > 0$ *and* $X_r^- > 0$ *simultaneously*. Proof follows by assuming the contrary for some r. Then from equation (12) $P_r - \lambda_r - b \sum_s \delta_s = 0$, and from equation (13), $-P_r + \delta_r - c \sum_s \delta_s = 0$. Adding these together gives $-\lambda_r + \delta_r - (b + c) \sum_s \delta_s = 0$, which implies $-\lambda_r + \delta_r - a \sum_s \delta_s > 0$, a direct contradiction of equation (14), proving the result.

The conclusion shows that firms have an incentive to merge. The maximum profit method of producing the entire output of the insurance industry requires that either purchases or sales be made in each state in the capital market, never both purchases and sales in one state. If there are two or more firms in the industry it is likely that one is selling in the capital market in some state for which another is buying in the capital market. Two such firms reduce costs by merging and eliminating the cross transactions.

Two brokerage firms could coexist with no incentive to merge only if they had identical patterns of trade in the capital market. Since there are no indivisibilities in the firm production function, it is technically possible that many firms could coexist, each with the same pattern of capital market use. The incentive to merge created by economies of internal clearing is akin to economies of scale but exists in the absence of technical scale economies.

Economies of internal clearing have another implication for equilibrium brokerage behavior. In the presence of such economies the vector ϕ in equation (14) is zero in all states in which there is trade, i.e., $W_r > 0$ implies $\phi_r = 0$. This is proved by observing that transactions costs must be paid in every state; that implies that either $w_r > 0$ or $X_r^- > 0$ for all states r. If $w_r > 0$ then $\phi_r = 0$, by the conditions of equation (14). If $w_r = 0$ then $X_r^- > 0$ and consequently $X_r^+ = 0$. Then the constraint $W_r - X_r^+ - w_r \geqslant 0$ is not binding and profit could be raised by selling the proceeds W_r in either the capital market or the Z-market, a contradiction of profit maximization. Consequently $W_r > 0$ implies $w_r > 0$ and thus $\phi_r = 0$, proving the result.

This implies that a fixed relationship exists between equilibrium prices in the W-market and Z-market. This relationship is

$$\lambda = \Phi(-a)\delta. \tag{17}$$

Now if $P^w = \lambda$ and $P^z = \delta$, as must hold in equilibrium, the consumer's budget constraint, from equation (9), becomes

$$\Phi(-a)W' - Z' = 0. \tag{18}$$

This shows that consumer satisfaction is independent of the absolute price level and depends only on the spread of prices between the W-market and Z-market. The price of one unit of goods with certainty in the W-market is $\sum_s \lambda_s$, and the price of one unit of consumption goods with certainty in the Z-market is $\sum_s \delta_s$. The relationship of these prices is, from equation (17),

$$\frac{\sum_s \lambda_s}{\sum_s \delta_s} = (1 - Sa), \tag{19}$$

which is also a constant and independent of the absolute price level. Thus, from a welfare standpoint and from a technical standpoint the relationship between W-market prices and Z-market prices is unvarying.

☐ **Prices for a single insurer.** The absolute level of prices is important, however, for the analysis of competition among various kinds of brokerages and for an understanding of the problems of insurance regulation. The clearest way to present the absolute prices is to study the prices of a single brokerage which carries out all transactions for the consumers, while pricing like a competitor. This brokerage is an insurance brokerage because it satisfies the barter constraint of equation (10). It must satisfy the barter constraint, because all consumers do so as equation (9) demonstrates.

The absolute level of prices depends upon the degree of self-sufficiency, which is defined here as the number of states in which the firm sells in the capital market. Renumber states so that $X_s{}^+ \geqslant 0$ for $s = 1, \ldots, r$, and $X_s{}^- > 0$ for $s = r + 1, \ldots, S$. Here r measures the degree of self-sufficiency. With these equilibrium conditions the first r components in equation (12) are equal to zero and the last $S - r$ components in equation (13) are equal to zero. The variable λ_r in equation (12) may be eliminated by substituting from equation (14). The result relates the offer prices of the insurer, δ^I, to prices in the capital market by

$$P = A\,\delta^I, \tag{20}$$

where

$$A = \begin{array}{c} \\ r \\ S - r \end{array} \begin{array}{cc} r & S - r \\ \multicolumn{2}{c}{\begin{pmatrix} \Phi(-a + b) & \Gamma(-a + b) \\ \Gamma(-c) & \Phi(-c) \end{pmatrix}} \end{array}. \tag{21}$$

The matrix A has been written as a decomposed matrix using familiar notation. The dimensions of the component matrices are indicated by notations at the borders of the matrix. Note that the off-diagonal components are rectangular matrices while the diagonal components are square matrices. The determinant of A is $\Delta(r)$, which is

$$\Delta(r) = 1 - Sc + r(b + c - a) \tag{22}$$

and the inverse is

$$A^{-1} = \frac{1}{\Delta(r)} \begin{array}{c} \begin{array}{cc} r & S - r \end{array} \\ \begin{pmatrix} \Phi(a - b) & \Gamma(a - b) \\ \Gamma(c) & \Phi(c) \end{pmatrix} \end{array} \begin{array}{c} \\ r \\ S - r. \end{array} \tag{23}$$

Then

$$\delta^I = A^{-1}P. \tag{24}$$

This may be written component-wise as

$$\delta_s{}^I = P_s + \frac{a-b}{\Delta(r)} \sum_{t=1}^{s} P_t \quad \text{for} \quad s \leqslant r \tag{25}$$

$$\delta_s{}^I = P_s + \frac{c}{\Delta(r)} \sum_{t=1}^{s} P_t \quad \text{for} \quad s > r. \tag{26}$$

When r increases from r_0 to $r_0 + 1$, each price $\delta_s{}^I$ derceases for $s < r_0 + 1$. Since $a < b + c$, $\delta_{r_0}{}^I + 1$ also decreases. The absolute level of Z-market prices declines as the degree of self-sufficiency increases. Relative prices are, of course, constrained by equation (17), that is,

$$\lambda^I = \Phi(-a)\delta^I. \tag{27}$$

☐ **Various kinds of competition.** So far all brokerage has been performed by one large insurance firm. Consider the potential competitors for this firm. Another insurer who entered would have a higher number of states in which $X_s{}^- > 0$. For this reason the absolute level of prices he could offer would be higher than those of the original insurer. The second insurer would be competitive from a consumer's viewpoint, however, because he can offer the same relationship between bid and asked prices. If consumers are constrained to satisfy a barter constraint in their trade with each insurance firm, then they are indifferent between the original insurance firm and the competitor. The two firms have an incentive to merge, but if they should merge, it would not leave them free from other competition. Another insurer could offer the same barter prices that they offer in spite of lacking their superior pattern of access to the capital market. On the basis of this analysis there is no reason why competition should not prevail among insurers.

If institutional or legal factors should make insurance a monopoly, it would still be constrained in the extent to which it could manipulate prices. Consumers would have the alternative of selling their contingent claims to incorporators and buying other contingent claims from reinsurers. Recall that incorporators are brokerages constrained to satisfy $Z = 0$ and reinsurers constrained to satisfy $W = 0$. Since there are economies of internal clearing, the transactions costs involved must be higher than those in the direct transfer of these contingent claims.

The reinsurer has $X_s{}^- > 0$ in every state. This corresponds to the case $r = 0$ in the previous section, and from equations (23) and (24) it follows that the Z-market price of the reinsurer, δ^R, is given by

$$\delta^R = \Phi\left(\frac{c}{1 - Sc}\right)P. \tag{28}$$

An incorporator has $X_s{}^+ > 0$ in many states, and it has shadow prices, δ^F, at which it could just afford to supply contingent claims in the Z-market, provided it were legally able. Because the incorporator has $X_s+ > 0$ in some states, $\delta^F < \delta^R$ by application of the result of the previous section. Since equation (14) is satisfied with $\phi = 0$, the incorporator's offer price in the W-market, δ^F, satisfies $\lambda^F = \Phi(-a)\delta^F$, and consequently

$$\lambda^F < \Phi(-a)\delta^R. \tag{29}$$

Comparing this to equation (27), it is evident that the reinsurance-incorporation route can never duplicate the costs and prices of competitive insurance. This possibility does, however, bracket the possible price spread offered by a monopoly insurer. The stringency of this check on potential monopoly depends upon the extent of economies of internal clearing. If economies of internal clearing are nonexistent, the reinsurance-incorporation route is as inexpensive as insurance.

5. Regulation

■ Among the many functions of insurance regulation is the governance of insurance prices. Abstracting from all other regulatory functions, consider the concerns of the regulator in setting insurance prices. The consumer is not affected by the absolute height of insurance prices P^w and P^z, but only by their relative values, as observed above. In any competitive equilibrium, prices satisfy $P^w = \Phi(-a)P^z$, from equation (17), and the consumer's constraint is equation (18), $\Phi(-a)W^j - Z^j = 0$. The constraint is independent of the level of prices. Considering the situation in this way, the insurance regulator ignores the absolute level of prices of contingent goods and in particular their capital market prices. An insurance contract between firm f and individual j is $(W^{f,j}, Z^{f,j})$. The first vector is the individual's sales to the insurance firm; the second vector is the firm's sales to the consumer. In equilibrium the prices P^w and P^z for individual contingent claims are disregarded, as are the capital market prices P. Insurance regulation consists of constraints on individual contracts which specify that given $Z^{f,j}$, *the* $W^{f,j}$ may not be increased above a certain level.

The difficulty with regulation stems from the possibility of capital market price shifts. Although capital market prices are irrelevant to the consumer and ignored by the regulator, they are central to the operations of the insurance firm. Initially at capital market prices P, the value of a contract to an insurance firm is $\pi^{f,j}$, where

$$\pi^{f,j} = \lambda \cdot W^{f,j} - \delta \cdot Z^{f,j}. \tag{30}$$

Because of equation (17) this may be rewritten as

$$\pi^{f,j} = \sum_s(\delta_s - a\sum_t \delta_t)W_s^{f,j} - \sum_s \delta_s Z_s^{f,j}. \tag{31}$$

Now consider a reevaluation of states in the capital market resulting in prices P'. This is assumed to leave unchanged the value of a unit of income with certainty, that is, $\sum_s P_s = \sum_s P'_s$. From equations (25) and (26) it follows that

$$\delta'_s - \delta_s = P'_s - P_s \tag{32}$$

$$\sum_s \delta'_s \sum \delta_s = 0. \tag{33}$$

The effect on the contract's value is the difference between the new value $\pi^{f,j'}$ and the old one, which reduces to

$$\pi^{f,j'} - \pi^{f,j} = \sum_s (P'_s - P_s)\, W_s^{f,j} - \sum_s (P'_s - P_s) Z_s^{f,j}. \quad (34)$$

If the insurance market is competitive, the initial value of the contract is zero. The reevaluation in the capital market may raise or lower the value of the contract to the company, and if the value drops below zero, the company wishes to raise $W^{f,j}$ or lower $Z^{f,j}$. In the presence of regulation that forbids raising $W^{f,j}$, given $Z^{f,j}$, or lowering $Z^{f,j}$, given $W^{f,j}$, the company wishes to drop the contract entirely. Thus the effect of binding regulation is to cause cancellation of some contracts.

The situation is little different if the insurance market is not competitive. In this case each $\pi^{f,j}$ is initially greater than zero. Binding price regulation may force a positive value down without making it negative, but if profits are low because of effective competition from reinsurers and incorporators, binding regulation may still result in the cancellation of contracts. The more nearly competitive the industry is, the greater the number of cancellations. The incidence of cancellations is predictable, being centered on contracts with $Z_s^{f,j}$ positive for states in which prices rise.

Assuming again that the insurance market is not entirely competitive, the case of profit regulation is similar. The means of regulating profit is the imposition of a proportional reduction in each entry of $W^{f,j}$, for all j. Assuming that random reevaluations have occurred in the capital market and that insurance price regulation has been binding in the past, some contracts have values near zero, while others yield a distinctively positive profit. The reduction in $W^{f,j}$ will force some contracts below the margin of profitability. In this case the incidence of cancellations is random.

☐ **The role of market imperfections.** These results depend upon an imperfection of insurance markets beyond the existence of transactions costs. Although every technical risk may be insured, the risk inherent in a reevaluation of capital market prices is not insured. Insuring the latter risk is theoretically possible because some individuals benefit from reevaluation in the capital market. The profitability of their contracts rises and competition among insurers tends to permit the lucky individuals to realize savings. In an insurance system that insured the capital market risk, these gainers would be transferring income to those injured by the price shift.

The goal of insurance price regulation is compensatory when it strives to reduce the cost which a capital market price shift imposes on specific individuals. The means to the goal is to place the contracts involved under binding constraints. Compensatory regulation aims to remedy the imperfection of insurance markets, but it fails because its impacts are not symmetrical. The gainers from a capital market price shift cannot be taxed to aid the losers because competitive forces tend to lower prices of their contracts, a change which is not constrained by regulation. The desired transfers can only be imposed by preventing price competition. A simpler and better strategy, which also promotes efficiency of risk-bearing generally, is to encourage contracts which insure the capital market risk and to preserve the competitive forces that exist.

References

1. Arrow, K. J. "Le rôle des valeurs boursières pour la répartition la meilleure des risques," in *Econometrie*, CNRS, Paris, 1953, pp. 41–47. Translated as "The Role of Securities in the Optimal Allocation of Risk Bearing," *Review of Economic Studies*, Vol. 31, No. 2 (April 1964), pp. 91–96.
2. Brainard, W. and Dolbear, F. T. "Social Risk and Financial Markets" *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 360–370.
3. Debreu, G. *Theory of Value.* New York: John Wiley & Sons, 1959.
4. Kihlstrom, R. and Pauly, M. "The Role of Insurance in the Allocation of Risk." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 371–379.
5. Marshall, J. M. "Theory of Insurance and Applications to Insurance Regulation." Unpublished Ph.D. dissertation, Massachusetts Institute of Technology, 1969.
6. Willett, A. H. *The Economic Theory of Risk and Insurance.* Homewood, Ill.: Irwin, 1959. First published as *Studies in History, Economics, and Public Law*, 14(2), New York, 1901.

The economics of product safety and imperfect information

Victor P. Goldberg

Associate Professor of Economics
University of California, Davis

In an earlier article in the Bell Journal, Walter Oi analyzed the economics of product safety. His primary emphasis was on a full information model. However, since product safety is only a problem if consumer information is inadequate, policy implications of a full information analysis are suspect. This paper stresses the importance of treating product bans and liability rules as part of the consumer's information processing system.

■ In the Spring 1973 issue of this journal, Walter Oi presents an economic analysis of the problem of product safety.[1] His analysis begins with a review of the recommendations of the National Commission on Product Safety (NCPS)[2] which proposed that stricter product liability (*caveat venditor*) be established and that a federal agency be empowered to establish minimum safety standards for existing products, banning those that fail to meet the standards.[3] Oi's concern is primarily with the latter issue, although he does treat the issue of consumer versus producer liability somewhat.

Oi's demonstration that a government ban of "unreasonably hazardous products" forces some consumers to lower indifference curves[4] is a complicated proof of an elementary notion: if one product is eliminated from the market by government fiat, social welfare is decreased. Fundamental to Oi's demonstration is the assumption that "all parties are fully informed."[5] Coupling this assumption with the primacy of consumer preferences, an emphasis on allocative efficiency rather than distribution,[6] and an implicit assumption that no externalities are involved, Oi reaches the conclusion that *caveat emptor* and no government prohibition are the most desirable policy. But, as Oi points out, the NCPS explicitly disavows the full information assumption:[7]

The author would like to thank the following for their comments on earlier drafts: Gerrit Buddingh, Rodolfo Gonzalez, Ronald Hunt, Blair Lord, Thomas Mayer, William Moss, Alan Olmstead, James Roumasset, and Elias Tuma.

[1] Oi [5]. For analyses which are not so rigorous as Oi's but which reach similar conclusions, see McKean [3] and Buchanan [1].

[2] In [4]. I am not attempting here to vindicate the NCPS recommendations; an evaluation of those standards is far beyond the scope of this paper.

[3] *Ibid.*, pp. 7–8.

[4] Oi [5], pp. 5–14.

[5] *Ibid.*, p. 5.

[6] Oi does briefly discuss the possible redistributive results of producer liability (pp. 15, 27), but generally the emphasis is on allocative efficiency.

[7] Oi [5], p. 23.

> According to the NCPS report, the nation is presently facing a serious product safety problem because consumers are unaware of the risks of using increasingly complex products, unable to cope with these risks even if they knew about them, unable or unwilling to get accurate product information, or misled by questionable advertising practices. In short, the consumer is a pigeon.

To discredit the NCPS report it is necessary for Oi to show either that his conclusions are valid given the NCPS assumption concerning information, or that his assumption of full information is an accurate representation of the world. The first cannot be demonstrated with the apparatus employed by Oi; and, as Oi acknowledges,[8] the full information assumption is not meant to portray the world accurately. When he does try to relax the full information assumption, he is forced to abandon the analytical rigor of his previous analysis and resort instead to a much more casual approach.

In this more casual analysis the author suggests three ways of dealing with the problem of product safety. One is *laissez faire* with complete insurance against product accidents; this he considers only partially successful because the private market is unlikely to provide optimal insurance.[9] Second, he argues that "to the extent that there are external benefits from the production of better information, there is a basis for intervention, possibly via the government subsidization of information production."[10] This appears to be his favorite approach, although this preference can only be inferred. The third approach is that of the NCPS to protect consumers "by establishing barriers to entry of unreasonably hazardous products."[11] Oi does not find this attractive, since setting standards and testing products are costly,[12] decisions will be made by a bureaucracy "akin to . . . the Food and Drug Administration,"[13] and such rules restrict the consumer's freedom of choice.[14]

Oi's policy preference for government production of information stems from a misreading of his initial (full information) analysis. Since the market worked well with full information, and since the world is characterized by incomplete information, then concluding that improving the quality of information will lead us closer to the desired results of the full information model has a reasonable ring, even though the conclusion does not follow logically. The call for more information is not a simple and operative command. How should the supply of information be increased? Should the government commission and publish regular product safety reports? Should it require that product labels clearly report all the relevant product safety information including the probability distributions of various sorts of product failure? Since the consumer is regularly bombarded with much more information than he can possibly process, he might easily fail to perceive product safety information unless it is presented with great *élan*. Should product safety information be disseminated with attention-grabbing television commercials (like the antismoking commercials of a few years back)? Or might we consider the erection of a centrally located tote board, perhaps somewhere in Kansas, which would give the current odds on exploding pop bottles, automobile accidents, and a myriad of other potential safety problems?

---

[8] *Ibid.*, p. 19.
[9] *Ibid.*, p. 24.
[10] *Ibid.*, p. 26.
[11] *Ibid.*, p. 26.
[12] *Ibid*, p. 26.
[13] *Ibid.*, p. 26.
[14] *Ibid.*, p. 27.

As the above examples make clear, the issue is more than simply how much and what kind of information ought to be *produced*. We must also concern ourselves with the very difficult problems of *distribution* of information. Enlarging the pool of product safety knowledge does not necessarily improve the welfare of the consumer, if the information is not accessible; and getting the information from the producer to the consumer is not a simple task. Information might be conveyed by banning certain products or by adopting specific liability rules. That is, product bans and liability rules should be treated as part of the information processing system. Let us elaborate on this point.

It is convenient to view the consumer as delegating part of the information processing task to an agent.[15] Consider the problem an agent faces in conveying to the consumer information that will permit him to balance the risks of automobile accidents against the benefits of automobile travel. Possibly he can present the consumer with a multidimensional array of expected damage with outcomes depending on the driver's skill, the condition of the car, the type of roadway, weather conditions, traffic conditions, and speed of the car. The expected damage would have to be translated by the agent into subjective damage estimates; that is, the agent would have to garner information on the consumer's preferences. It should go without saying that the costs of such service would be prohibitive and that, even if the agent could satisfactorily provide such information to the consumer, it would be too complex for him to apply. (He would have to hire another agent to reduce the data to manageable proportions.)[16] Efficient performance by the agent would almost certainly require that he aggregate the individual consumers into one or a few representative consumers to simplify the problem of estimating subjective damage and collapse the information developed into a few simple rules so that the consumer could actually use it.

The rules can take the form of published information or public exhortation ("buckle up for safety"). Or they may take the form of specific prohibitions or requirements.[17] It is not at all unreasonable to expect that the best system of conveying information would entail some combination of enforceable (but not necessarily completely enforced) minimum driver standards, minimum car standards, and maximum speed limits.[18] By the full information model a proscription of the good—"a car trip in California at a speed in excess of 70 mph"—is clearly inefficient; but in a world of imperfect information this can be an extremely effective way of collapsing a large array of information into a very simple rule. Clearly, there will exist individuals who, given their risk preferences, driving skills, the condition of their car, etc., would be better off if they could legally drive at speeds in excess of 70 mph; conversely, there will be others who are a menace to themselves and society at speeds of 40 mph. But given the economies of interpersonal spreading of information costs inherent in this problem, such results are inevitable.

---

[15] The agent can be other consumers (word-of-mouth), producers, retailers (advertising, brand names, etc.), or government. For greater discussion on the role of the agent, see Goldberg [2], pp. 2–8.

[16] In Oi's full information model all this is done costlessly.

[17] Prohibitions and requirements are two sides of the same coin. Thus, the prohibition of items with safety problem X is the same as the requirement that products do not have problem X.

[18] One possible economizer on information which seems to get minimal

As a second example, consider the problem of exploding soda bottles.[19] There is a finite probability that a soda bottle will explode and injure a consumer (or third party) and this probability depends in part on the construction of the bottle and its contents (the amount of carbonation); it also depends in part on the actions of consumers. Oi's full information model would suggest that consumers should have the right to choose cheaper, higher risk bottles if they so desire. But without full information, the free choice solution loses much of its appeal. How is the consumer supposed to ascertain the probabilities of different soda bottles' exploding? By experience? Oi suggests that the "consumer who regularly buys the risky product will eventually learn" that he will get a certain percentage of risky products.[20] But accidents of this sort are sufficiently rare that learning by doing will be an impractical solution.[21] Nor does provision of estimates of bottle failure by the government or private producers appear to be a particularly appealing approach. Generating the data and keeping the information current are both likely to be complicated tasks, given the thousands of different bottling plants in the country. And the distribution problem—providing the data to the consumer at the time of need[22]—is likely to be even more costly and complicated.

Some restrictions concerning "clearly" unsafe bottles might be useful, but the real problem concerns bottles that are not clearly unsafe. Suppose that we were to adopt a policy of penalizing the producer for product failure. This will have little direct effect on the consumer's behavior, since he will find it too costly to attain information on bottle integrity both before and after such a penalty

---

support is a requirement that cars be constructed so that they cannot exceed a certain speed.

[19] See *Escola* v. *Coca-Cola Bottling Co.*, 24 Cal. 2d 433, 150 P.2d 436 (1944).

[20] Oi [5], p. 7. It should be noted that this statement is made ir Oi's discussion of the full information model.

[21] Learning from one's own experience is even more impractical if the injury is a very serious one. In the extreme case of a fatal accident, of course, the learning experience might be profound, but the learning curve is abruptly truncated.

[22] The problem is even further compounded since bottle explosions can injure not only soda drinkers but other shoppers who have no desire to purchase soda, but who must pass soda bottle displays in their tour of the supermarket. While one might conceive of avid soda drinkers' attempting to gain information on bottle safety (although this is dubious), it is extremely difficult to conceive of the nonsoda buying shoppers' seeking such information.

system is instituted. But the producer will be able to spread the information costs over a large enough number of transactions to ascertain whether the penalty is of sufficient magnitude to induce him to adopt a safer bottle or to invest in research and development to develop a safer bottle. The penalty can take the form of a fine[23] for producing bottles that do not meet certain quality standards, a fine to be paid only for actual product failure or liability for damage resulting from product failure (*caveat venditor*), etc. The point is that the consumer's information processing will be grossly inadequate if he is left to his own devices, but the task can be considerably eased by collective action—penalizing the producer. As in the previous example, the collectively established decision rules will represent the agent's perception of the representative consumer's preferences, barring some consumers from taking risks that they would be willing to accept and subjecting other consumers to risks that they would be unwilling to accept.

What conclusions does this discussion suggest? Given that product safety is a problem only if consumer information is inadequate, an analysis of the problem which begins with the assumption that information is perfect (and which cannot be extended to relax this assumption) is likely to yield suspect policy implications;[24] the way in which alternative resource allocation mechanisms produce and transmit product safety information is the ultimate concern, and this Oi must assume away. Since information is important, economists are going to have to get their hands dirty and investigate the characteristics of alternative means of producing and distributing product information including product bans and liability rules.

In a world with imperfect consumer information concerning the risks of product failure (that is, the world in which we live), public policy makers face a difficult tradeoff. Some individuals will correctly perceive the risk of a particular activity, decide that they are willing to face them, and yet be prevented from doing so by the law. Conversely, others will incorrectly perceive the risks and will end up exposed to risks they would otherwise be unwilling to take. Increasing the "freedom of choice" of the former group will, in many instances, entail an increased exposure to undesired risks for the latter. Oi implicitly is arguing that the number of people in the second category will generally be small.[25] I am suggesting that the relative size of the two groups is open to question and that for many problems of product safety Oi is overstating the membership of the first group. Thus, while strict product liability will tend to deprive some consumers of the choice of cheaper but more dangerous bottles, it will also tend to protect another group from facing a greater risk of bottle explosion than they would have expected in a world of *caveat emptor*. While there is not hard evidence to back it up, my guess is that in this instance virtually everyone is in the second group and that the strict liability would increase virtually everyone's welfare.

---

[23] Or a jail sentence might be substituted for a fine.

[24] The perfect information assumption might still be useful for certain positive purposes, however.

[25] Oi might also be interpreted as arguing that the first group's preferences *should* count for more; "freedom of choice" is an end in itself in Oi's value system (and in the value system of many other economists as well).

References

1. Buchanan, J. M. "In Defense of *Caveat Emptor*." *University of Chicago Law Review*, Vol. 38 (Fall 1970), pp. 64–73.
2. Goldberg, V. P. "Consumer Choice, Imperfect Information, and Public Policy." IGA Research Report No. 26. Davis, Calif.: University of California, Institute of Governmental Affairs, August 1973.
3. McKean, R. N. "Products Liability: Trends and Implications." *University of Chicago Law Review*, Vol. 38 (Fall 1970), pp. 3–63.
4. National Commission on Product Safety. *Final Report of the National Commission on Product Safety*. Washington, D. C.: U. S. Government Printing Office, June 1970.
5. Oi, W. Y. "The Economics of Product Safety." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 3–28.

The economics of product safety: a rejoinder

Walter Y. Oi

Professor of Applied Economics
Graduate School of Management
University of Rochester
and Visiting Senior Research Economist
Industrial Relations Section
Princeton University

The model of the demand for a risky product (in the Spring 1973 issue of this Journal), has been criticized because it assumed perfect information. Goldberg contends that the implications of that model are wholly irrelevant for a world of imperfectly informed consumers. He further asserts that the model is incapable of relaxing the full information assumption. In this rejoinder, I first restate the salient features of the earlier model. The minimization of the sum of accident costs and accident prevention costs is surely the appropriate goal for public policy toward product safety, but this goal is largely ignored in Goldberg's comment. Contrary to his assertion, I show my model can easily be modified to analyze the implications of imperfect information within the context of product risks that can be described by one parameter. Some remarks are also offered on the costs and benefits of product bans as well as on directions for further research.

■ Public policy with respect to product safety should strive to minimize the *sum* of accident costs and accident prevention costs. Injuries resulting from product-related accidents can entail costs in the form of medical bills, wage losses, litigation expenses, or psychic costs. Accident prevention costs must include both the higher costs of safer products and the monetary value of any utility losses arising out of legal restrictions that preclude the sale of cheaper but more hazardous products. This principle is surely the right one for evaluating the merits of public policy proposals. In order to apply this principle, one needs an analytic model of the market for risky products, and I tried to develop such a model in my earlier paper. After emphasizing the imperfect information assumption of the National Commission on Product Safety (NCPS) report, Goldberg questions the relevancy of my analysis with the contention, "To discredit the NCPS report it is necessary for Oi to show either that his conclusions are valid given the NCPS assumption concerning information, or that his assumption of full information is an accurate representation of the world."[1] Before commenting on Goldberg's criticisms, it may be helpful briefly to review the essential features of my model.

---

[1] In [1], p. 684.

The theoretical model developed in Section 2 of my paper explicitly invoked two simplifying assumptions: (1) the product risk was inherent and could not be altered by accident-avoiding actions, and (2) consumers and producers were fully informed about these inherent product risks. I also tacitly assumed that the products do not inflict damages on third parties and that the risks were unrelated to other attributes of the product. The model is based on the premise that the consumer who buys a risky product is truly demanding a joint product consisting of the utility generated by nondefective units and the disutility (or accident cost W) suffered in the event of a defective unit. The *full price* P^* of a risky product can thus be expressed as the sum of the average price of a nondefective unit, $\hat{P} = P/\pi$, plus the expected accident cost from a defective unit, $W\left(\frac{1-\pi}{\pi}\right) = W\gamma$, where π is the proportion of nondefective, good units. If consumer liability prevails in this ideal world of inherent risks and full information, each individual maximizes utility by demanding that product grade which yields the lowest full price, $P^*_j = \hat{P}_j + W\gamma_j$. Consumers who suffer low accident costs W in the event of a defective unit will demand the more hazardous product grade because they are willing to incur higher expected accident costs in order to obtain a lower average price $\hat{P}$ for nondefective units.

A shift to strict producer liability (*caveat venditor*), substantially alters the situation. Under producer liability, the seller of a risky product is compelled to offer buyers a tied package consisting of the parent risky product plus a full coverage insurance policy against all accident costs. Since consumers are now fully compensated for all accident costs, they will now maximize utility by demanding that product grade with the lowest full supply price $\bar{P}_j$. However, under *caveat venditor*, the full supply price must include a risk component equal to the average damage cost per defective unit, $\bar{W}\gamma$, applicable to *all* consumers. The risk component, $\bar{W}\gamma_j$, in the full supply price will obviously be larger for riskier products with lower values of π_j. This fact apparently prompted earlier writers to conclude that a shift to strict producer liability would discourage the production of riskier products. However, a careful analysis reveals that under some conditions, this conclusion is wrong. A correct comparison should contrast the structure of full supply prices $\bar{P}_j$ under producer liability with the structure of full prices P^*_j that buyers would have faced under consumer liability. It was shown in my earlier paper that when the average damage cost $\bar{W}$ for all consumers was less than the critical damage cost W^* incurred by the marginal consumer, a shift to producer liability in a world of fully informed buyers would lead to the disappearance of the safer product grade, which, in turn, implies higher accident costs. Moreover, a shift to producer liability leads to a redistribution of income away from consumers who incurred low accident costs (presumably poorer consumers) to the richer consumers who experienced high accident costs. These implications could not have been derived without the aid of a formal analytic model.

The theoretical model and its assumptions were *not* meant to give us an accurate description of the real world, but this is as it should be. Theory must abstract from many real-world complications in order to direct attention to the structural equations of the economic process under investigation. The theory of consumer demand is not invalidated by observing that my purchases, and probably yours, do not satisfy *all* of the axioms of revealed preference.

Let me turn now to Goldberg's criticisms of my paper. His first example of auto safety standards is mainly concerned with a set of risky products that inflict damages to third parties as well as to the buyer. I admit that my analysis neglects these externalities and is limited to product risks involving only the buyer and seller. His second example of the exploding soda pop bottle is a favorite example in both the popular and legal literature on products liability. The NCPS even commissioned a study to the Cornell Aeronautical Laboratory to analyze the problem and to estimate the added production cost for a bottle with greater integrity. Their findings were not surprising. The heavier bottles for which we must leave deposits are far less likely to explode than the lighter throwaway bottles. Further, the safety of the heavier deposit bottles declines steadily with reuse. The probability of an explosion also depends on internal carbonation pressure, external air pressure and temperature, and the handling (banging around) of bottles by both the sellers and consumers. Although the study did not compare deposit bottles to cans, I am prepared to wager that can integrity will be vastly superior to bottle integrity, even for the proposed superior design which would have added roughly one cent to the cost of a bottle. I am not aware of any evidence (and Goldberg is also silent on the subject) about the proportion of exploding bottles that inflict damages to third parties. If this proportion is small, we can ignore the externality issues in discussing the bottle example.

Goldberg's principal objections are concerned with the issue of imperfect information which, in his opinion, was poorly analyzed in my paper. If I may paraphrase him, my analysis suffered because of three "difficulties":

1. The conclusion that improving the *quality* of information will lead us closer to the desired results of the full information model has a reasonable ring even though the conclusion does not follow logically.
2. Attention is directed only to the production of information, and the dissemination of information is totally neglected.
3. The implicit definition of information systems in the Oi paper is unduly narrow and fails to recognize that product bans and liability assignment are part of the information processing system.

In the context of my model, a narrow definition of information is sufficient since the product risk is described by a single parameter π, the proportion of good units. I fully agree that "information" is not a homogeneous good like bunker C fuel oil and that the information requirements for some choices can be extremely complex. The agency relationship is incorporated into many of our institutions. I rely on my doctor to prescribe medication and on my contractor to choose safe qualities of wood and glass for the new family room. "Information" is an elusive concept, and most of us (including Goldberg and myself) have shied away from writing down an operational definition of the concept. I simply do not understand his concept of an "information processing system" or how product bans and legal assignment of liability are part of that system. With respect to the second alleged "difficulty," information that is potentially available but inaccessible to consumers is like wheat in the fields that cannot be shipped to market. Neither the information nor the wheat has been "produced" in an economic sense.

The charge that I contend is wholly unwarranted is contained in his concluding comments: "Given that product safety is a problem only if consumer information is inadequate, an analysis of the problem which begins with the assumption that information is perfect (and which cannot be extended to relax this assumption) is likely to yield suspect policy implications. . . ."[2] Product safety will be a public policy issue as long as policy-makers believe that even fully informed consumers must be prevented from exposing themselves to risks which the policy-makers regard as socially unacceptable. The prohibitions against flammable fabrics, *even if* the fabrics were clearly labelled, exemplifies this point. More importantly, my model can easily be extended to relax the full information assumption.

In the context of my model in which the product risk is described by one parameter π, the proportion of good units, imperfect information can only mean that the individual consumer errs in his subjective estimate of π. To facilitate the exposition, let Ψ_j denote the individual's subjective estimate of π_j, while Q^*_j is the perceived full price of the risky product given his possibly incorrect estimate Ψ_j. Thus,

$$Q^*_j = \frac{P_j}{\Psi_j} + W\left(\frac{1-\Psi_j}{\Psi_j}\right).$$

Given his possibly erroneous estimates of the product risks, the full prices Q^*_j may differ from the true full prices P^*_j. However, the individual consumer would still maximize his expected utility by demanding that product grade which minimizes his estimate of the full price, namely Q^*_j. He will thus demand the safer product X_1 when his assessment of the damage cost in the event of a defective unit W exceeds his estimate of the critical damage cost W^o; that is when,

$$W > \frac{\Psi_2 P_1 - \Psi_1 P_2}{\Psi_1 - \Psi_2} = W^c. \tag{1}$$

With perfect information, $(\Psi_j = \pi_j)$, this individual would have demanded the safer product if the following inequality were satisfied:

$$W > \frac{\pi_2 P_1 - \pi_1 P_2}{\pi_1 - \pi_2} = W^*. \tag{2}$$

We may say that the individual underestimates the product risk when $\Psi_j > \pi_j$, meaning that he overestimates the proportion of good units. It is alleged that imperfectly informed consumers will typically underestimate product risks.[3] With the aid of our model, we can analyze how such imperfect information—which manifests itself in a discrepancy between the true and perceived proportions of good units, $\pi_j < \Psi_j$—will affect (1) the individual's choice between two grades of a risky product and (2) his demand for units of the parent risky product. If he overestimates the proportion of good units by the same *proportion* for both grades of the risky product, $\Psi_j = (1 + \delta)\pi_j$, then it is apparent from equations (1) and (2) that $W^c = W^*$. In this special case of proportional errors in estimating product risks, if he chose to demand the riskier product with imperfect information—meaning that the inequality in (1) was reversed—he would still choose the riskier product when he is provided with accurate information about the risks, $\Psi_j = \pi_j$. If, however, individuals overestimate the proportion of good units by the same *nominal* amount, $\Psi_j = \pi_j + \Delta$, we have,

---

[2] In [1], p. 687.

$$W^c = W^* + \Delta \left(\frac{P_1 - P_2}{\pi_1 - \pi_2} \right).$$

Since $P_1 > P_2$, it follows that $W^c > W^*$. In this event some misinformed consumers could have been misled into demanding the riskier product X_2, because the perceived full price of X_2 would have been lower, $Q^*_2 < Q^*_1$, *given* their erroneous estimates of the product risks. Imperfect information is not enough to guarantee that consumers will shift their choices in favor of the riskier product. The imperfect information must be biased in such a way that the underestimation of product risks is greater for the more hazardous product grade; i.e. the ratio of the true to perceived chances of getting a defective unit from the riskier product X_2 must exceed that for the safer product X_1:

$$\left(\frac{1 - \pi_2}{1 - \Psi_2} \right) > \left(\frac{1 - \pi_1}{1 - \Psi_1} \right).$$

We can further analyze how an underestimation of product risks will affect the demand for units of the parent risky product. An underestimation will reduce the perceived full price of the risky product, $Q^*_j < P^*_j$, which in turn, expands the demand for good units Z. However, as I pointed out in equation (12b) of my earlier paper,[4] a rise in the proportion of good units π (whether real or imagined) can be accompanied by either a rise or a fall in the demand for the risky product X_j. If the risk elasticity of demand is negative, $\eta_\pi < 0$, imperfect information which underestimates the risks will lead to a decrease in the demand for X_j, thereby lowering aggregate accident costs. In my earlier paper, I did not elaborate on this extension of the model which relaxes the assumption of full information, because I felt that it was straightforward.

---

[3] To the best of my knowledge, no one has convincingly documented this allegation. Casual evidence is unreliable and tends to be biased. As Schelling [5] has pointed out, we often know when someone dies for a lack of safety, but we are often unaware of the number of lives that are saved because of safety. It could easily happen that risk averse consumers would overestimate product risks. In this event, we would get results that are opposite to those developed in the text.

[4] In [4].

With respect to my first "difficulty," it is surely true that in the context of my model, if consumers had the *right* information (implying that $\Psi_j = \pi_j$ for all product grades), consumer liability and utility maximization would result in an optimum allocation of resources which minimizes the sum of accident and accident prevention costs. Goldberg's criticism makes sense to me only if an improvement in the *quality* of information does *not* reduce the magnitude of the differential between true and perceived product risks, $(\pi_j - \Psi_j)$. The implications of my analysis might be turned around if the risky products in the model were described in a more complicated attribute-space, but Goldberg has not done this. An interesting case arises when the risks are technically related to other product attributes. One could purchase a safer and less caustic dishwashing detergent, but he might have to settle for dirtier dishes. One of the proposed "safer designs" for rotary lawnmowers involved an extremely slow speed for the rotary blade; very little was said about the grass-cutting efficiency of that design.

Finally, in evaluating the merits of a proposed product ban, the economist should be concerned with the measurement of the costs and benefits of the proposed product ban. Identifying "two groups of consumers" only gives us a head count and should be viewed only as a starting point. A product ban on throwaway bottles is nothing more than the imposition of a discriminatory prohibitive tax. There is, in principle, some structure of bottle taxes such that given these taxes, the utility-maximizing consumption bundle for each consumer would contain *no* throwaway bottles. In a world of fully informed consumers, the economic cost of this bottle tax (ban) can be measured by the lump-sum income transfer that consumers would be willing to pay to eliminate this tax.[5] One benefit of the ban which is difficult to measure is the value of preventing fully informed consumers from exposing themselves to what the policy-makers (or society) regard as a socially unacceptable risk. I have briefly discussed this; even if consumers are willing to buy clearly labeled flammable fabrics, our laws prevent them from exposing themselves and their children to these socially unacceptable risks. The principal benefit of a product ban derives from the presumption that consumer expectations about the expected damage costs (resulting from buying unreasonably hazardous products) are systematically *below* the actuarial expectations. As a result, these imperfectly informed individuals consume "too many" unreasonably risky products which inflict inordinately high accident costs. In principle, if these consumers could have obtained the right information, they would *not* have demanded the riskier products. If the imperfect information presumption is accepted, the economist must still measure the benefits taking care not to count the costs that fully informed consumers would incur from a ban preventing them from buying riskier but cheaper products.[6] These benefits must exceed the costs to warrant an endorsement of the ban. Even after I learned something about the chances of being hurt by an exploding bottle—and observed at least half a dozen minor explosions—I still buy my soft drinks in dangerous throwaway bottles. In addition to measuring the benefits and costs, one must demonstrate that the costs of informing consumers exceed the net costs of a product ban. The NCPS view (which Goldberg seems to accept) is that the costs of effectively informing consumers are astronomically high; this view is almost entirely based on judgment on the part of the Commission members. Finally, McKean[7] has directed attention to a number of undesirable side-effects that could accompany an expanded program of product bans and restrictions; the most important of these, in my opinion, is the retarding impact on technological change and product innovations that would follow an expanded policy of strict standards and bans. The product safety problem is an important one, but I hope that future research will give us more empirical evidence and a richer theoretical model.

---

[5] The theoretical rationale for this measure of the loss in consumer surplus was developed by Hotelling [2]. In applying this measure, one should remember that when imperfect information is replaced by full information, it need not affect the consumer's optimal choice of product grade. Put in another way, when imperfectly informed buyers are told about the true risks, they may still choose to demand the riskier but cheaper products, and a ban would generate costs for them.

References

1. GOLDBERG, V. "The Economics of Product Safety and Imperfect Information." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 2 (Autumn 1974), pp. 683–688.
2. HOTELLING, H. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." *Econometrica*, Vol. 6, No. 1 (July 1938), pp. 242–269.
3. McKEAN, R. N. "Property Rights, Regulation of Chemicals, and Information Production." Unpublished manuscript, University of Virginia, 1972.
4. OI, W. Y. "The Economics of Product Safety." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 3–28.
5. SCHELLING, T. C. "The Life You Save May be Your Own," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington, D. C.: The Brookings Institution, 1968, pp. 127–176, Comments by M. J. Bailey and G. Fromm.

---

[6] In measuring these benefits, a careful distinction must be made between uncertainty and imperfect information. An individual's assessment of the accident costs after the explosion may be very different from his *ex ante* valuations on which he based his choice. There are some similarities between product risks and financial risks. When one-eyed widows are bankrupted by an 80-percent fall in the price of Disney Productions, was this loss the result of uncertainty or poor information? Would the welfare of the small investor be improved by a policy that prevented him from investing in highly speculative stocks?

[7] In [3].

On "duration" and the optimal maturity structure of the balance sheet

M. A. Grove

Professor of Economics
Economics Department and Center for Capital Market Research
University of Oregon

The concept of the duration of an income stream has been employed in a variety of contexts in applied economics and finance. In this paper, duration is used to consider the general problem of choosing the optimal structure of the balance sheet under conditions of risk. Though this problem has been examined elsewhere in the literature, it has usually been discussed in terms of ad hoc rules such as "immunization" or hedging. The basis for the discussion of the problem here is a formal model of portfolio choice developed in the framework of the expected utility hypothesis of investment behavior. An examination of the comparative statics of the model and a comparison of it with the immunization rule constitute the main body of the paper.

1. Introduction

■ The concept of the duration of an income stream has a long history in the literature of monetary theory and finance.[1] Macaulay,[2] to whom the concept is generally attributed, proposed it as the most appropriate measure of the maturity dimension of coupon bonds. Some implications of replacing "nominal term to maturity" by "duration" for the theory of the term structure of interest rates were explored by Hickman,[3] who also pointed out some applications of the measure to a number of problems involving the mathematics of bond prices. More recently, Hopewell and Kaufman[4] employ the measure in their reformulation of the traditional statement of the relationship between the volatility of bond prices and the nominal term to maturity.

The first precise statement of the relevance of duration to the more general problem of planning the maturity structure of the balance sheet was provided by Hicks in his *Value and Capital.*[5] There duration (or "average period" as Hicks calls it) appears as the elasticity of the present value of a stream of payments with respect to the discount rate employed to calculate the present

---

I am indebted to my colleague G. O. Bierwag and to a referee for helpful comments and suggestions. Responsibility for errors is mine.

[1] The literature on the duration concept has recently been surveyed by Weil [11].

[2] In [7].

[3] In [4].

[4] In [6].

[5] In [5].

value.[6] Hicks then shows that the relative appreciation or depreciation in the present values of a pair of payment streams which is induced by a change in their common discount rate depends on the relative duration of the streams. One need only take one of the streams to be a stream of assets and the other to be a stream of liabilities to deduce the Duration Theorem stated independently by Samuelson:[7] An increase (decrease) in interest rates will increase net worth if the weighted duration of the liability stream is greater than (less than) the weighted duration of the asset stream. If the appropriately weighted durations of these streams are equal, then net worth will be unaffected by small changes in interest rates.[8]

The significance of the Hicks-Samuelson Duration Theorem to the problem of planning the maturity structure of the balance sheet is then apparent: persons or organizations wishing to speculate on interest rate movements should adjust their asset and liability streams so that these streams have different weighted durations; persons or organizations wishing to avoid the risk of interest rate movements should adjust their asset and liability streams so that these streams have the same weighted durations. But in order to incorporate the Duration Theorem in a model of the maturity structure of the balance sheet, one must be prepared to state just how such adjustments in asset and liability streams are made and to specify the conditions underlying a decision by a person or organization to speculate or to hedge.

The "immunization rule" proposed by Redington in the actuarial literature[9] is one such effort in this direction which has received attention in some recent work by Fisher and Weil[10] among others. Under it, the decision maker always chooses equal values of the weighted durations of his asset and liability streams, i.e., he always acts to hedge his net worth against interest rate movements. But the rule seems to me to be far too narrow and mechanical to survive serious examination. Attitudes toward risk, tastes in wealth, and expectations play no role in it, and the circumstances under which departures from a hedged balance sheet may be optimal are left unexplained.

The purpose of this paper is to reformulate the problem in the framework suggested by the modern theory of portfolio choice. The risk problem is examined in Section 2. A model of the maturity structure of the balance sheet based on the Duration Theorem is then developed and discussed in Section 3. Some concluding notes are contained in Section 4.

---

[6] If x_t is the dollar amount of a payment due at date $t = 1, 2, \ldots, T$, and $b = 1/(1 + r)$ is the rate of discount defined for the interest rate $r > -1$, then $V(x) = \Sigma x_t b^t$ is the present value of the stream of payments $\{x_t\}$. The elasticity of V with respect to b is $m(x) = \Sigma t x_t b^t / \Sigma x_t b^t$, i.e., the weighted average date to maturity or duration of the stream $\{x_t\}$.

[7] In [9].

[8] To prove the Duration Theorem, let $\{A_t\}$ and $\{L_t\}$ be asset and liability streams, respectively, whose present values calculated at the common discount rate b are $V(A)$ and $V(L)$. Net worth is then $V = V(A) - V(L)$. If $m(A)$ and $m(L)$ are the durations of these streams, then $\partial V / \partial r = b[V(L)m(L) - V(A)m(A)]$, and the statement in the text follows at once.

[9] See [8]. It is only fair to state that this special case of the Duration Theorem was developed independently by Redington and refined by a number of other actuaries. A very clear and precise statement of the rule is contained in the paper by Wallas [10].

[10] In [2].

2. The risk problem

■ Imagine an investor with a definite investment horizon planning the current structure of his balance sheet. Suppose that due to decisions implemented in the past, his cash inflow from interest payments and maturing bonds will be $A(t)$ dollars. Suppose further that his cash outflow in settlement of the obligations he has incurred will be $L(t)$ dollars on the same date. Let $t = 0$ be the current planning date, and $t = T$ be the investor's horizon. If we let t be a continuous variable on the closed interval $[0, T]$, then $A(t)$ and $L(t)$ are, respectively, the investor's asset and liability streams over his investment horizon.

The investor's wealth (net worth) is obtained by discounting these flows to the present under appropriate interest rate assumptions. Let $R(t)$ be the market rate of interest currently quoted on loans maturing at date t in the future. When t is a discrete variable, the structure of yields must be one for which the "long rate," $R(t)$, is the geometric average of current and one-period forward "short rates," $r(t)$, i.e., $R(t) = \{[1 + r(1)]\ [1 + r(2)] \ . \ . \ . \ [1 + r(t)]\}^{1/t} - 1$. When t is a continuous variable and interest is compounded continuously, the relationship between the instantaneous long rate, $R(t)$, and the intervening instantaneous forward short rates, $r(t)$, is given by the formulae

$$\begin{cases} R(t) = r(t) = \text{constant}, & \text{for} \quad t = 0 \\ R(t) = [\int_0^t r(x)dx]/t, & \text{for} \quad t > 0. \end{cases} \tag{1}$$

I shall refer to the function $R(t)$ as the market "yield curve" spanning the time interval $[0, T]$. The investor's wealth under the assumption that $R(t)$ is the yield curve is then given by the expression

$$\begin{aligned} W &= \int_0^T [A(t) - L(t)]\, exp\, [-R(t) \cdot t]dt \\ &= A - L, \end{aligned} \tag{2}$$

where A and L are present values of the investor's asset and liability streams, respectively. In the discussion to follow, I shall assume that both L and W are positive so that $A > L > 0$.

The investor faces two types of risk in his planning process. The first is that at some date in the future, his asset and liability streams will be different from what he expected. This kind of risk is present when debtors can default on their obligations and when creditors can revise payment schedules on debts incurred. I shall assume away this kind of risk in the discussion that follows.

The second kind of risk is present when interest rates can change. Wealth is the *difference* between the present values of the investor's asset and liability streams, and changes in interest rates can affect these values differently.

To see why, assume that $R(t) = R$ in (2) and differentiate W with respect to R. If the definitions

$$\begin{cases} m_A = (1/A) \int_0^T tA(t) \exp\,[-R(t)t]\, dt \\ m_L = (1/L) \int_0^T tL(t) \exp\,[-R(t)t]\, dt \end{cases} \tag{3}$$

of the durations of these two streams are employed in the resulting expression, the Hicks-Samuelson Theorem follows at once: the sign of the derivative dW/dr is the sign of the difference $Lm_L - Am_A$.[11]

The Hicks-Samuelson Duration Theorem suggests that m_A and m_L play a critical role in planning the current structure of the balance sheet. To develop the analysis further, assume that current interest rates are given by the yield curve in (1). For simplicity, I shall also make the following assumption about the investor's expectations: the investor is perfectly certain that the *only* yield curves that will materialize in the future are parallel displacements of the current yield curve. More formally, if h is a parameter independent of t, then the investor believes with certainty that any future yield curve will be of the form $R(t) + h$, for $0 \leqslant t \leqslant T$. I shall later assume that the investor is uncertain about the sign and magnitude of the parameter h.

Now let us calculate the investor's wealth under the assumption that the instant after he makes his plan the yield curve will be $R(t) + h$.[12] His wealth an instant after he makes his plan will then be

$$\begin{aligned} w &= \int_0^T [A(t) - L(t)] \exp\,\{-[R(t) + h]t\}\, dt \\ &= \int_0^T [A(t) - L(t)] \exp\,[-R(t) \cdot t] \exp\,(-ht)\, dt. \end{aligned} \tag{4}$$

If we replace $\exp(-ht)$ in (4) by the approximation $(1 - ht)$, then w is approximately

$$w = W + h(Lm_L - Am_A), \tag{5}$$

and the importance of the durations of the asset and liability streams is apparent.

Several possibilities are suggested by (5). If m_A and m_L can be adjusted by the investor, choosing them so that $Lm_L = Am_A$ hedges or immunizes W against small changes in interest rates.[13] An investor can also speculate on interest rate movements. If his best guess is that interest rates will rise, he can gamble on this guess by choosing $Lm_L - Am_A > 0$ and increase his wealth if his guess turns out to be correct. Similarly, he can gamble on a guess that interest rates will fall by choosing $Lm_L - Am_A < 0$.

---

[11] Cp., note 6, above.

[12] I shall add this assumption to avoid introducing another dimension of risk—the risk associated with *when* interest rates will change. To be sure, the timing of interest rate changes is an important consideration to investors in the "real world." But introducing it here destroys the static simplicity of the model. It may be convenient to think of the assumption in the text as follows: Let the date $t = 0$ be the instant following the investor's decisions and "the spinning of the wheel" that determines h.

[13] The approximation afforded by $1 - ht$ is a reasonable one only if h is small. Should h be large, retaining additional terms of the Taylor series is essential. In [3] I show that m_A and m_L are but *one* set of moments or "shape factors" that are important in immunizing a balance sheet. When h is large, an investor can improve his hedge by adjusting higher order moments in addition to m_A and m_L.

We can make these arguments more precise by assuming that the investor views the shift parameter h as a random variable with subjective probability density

$$f(h) \geqq 0, \qquad a \leqq h \leqq b. \tag{6}$$

Then the expected value of h,

$$E(h) = \int_a^b h\, f(h)dh, \tag{7}$$

is one form of a guess based on his expectations, and

$$E(w) = W + (Lm_L - Am_A)E(h) \tag{8}$$

employs this guess in a forecast of his wealth. Provided that wealth is desirable, speculating on interest rate movements then involves choosing m_A and m_L so that $Lm_L - Am_A$ has the same sign as $E(h)$. But the decision rule this seems to suggest, choosing m_A and m_L to maximize $E(w)$, is hardly an improvement over the immunization rule. Should $E(h)$ be different from zero, only the imposition of arbitrary constraints would prevent the investor from choosing $Lm_L - Am_A$ so that it has the same sign as $E(h)$ and is indefinitely large. I shall propose a more reasonable rule in the section to follow.

3. The model

■ In Section 2, I have suggested that the durations of the asset and liability streams can be regarded as decision variables for an investor planning the current structure of the balance sheet. In the model to be developed below. I shall make explicit use of these measures as decision variables in a model in which the investor faces interest rate risk. But before considering the technical features of this model, it is important to recognize that there are a number of different assumptions one can make that will govern the extent to which the durations of these streams can be adjusted by the investor. Since it is important to be clear about which assumptions are being made, I shall discuss them in detail in the paragraphs to follow.

One such set of assumptions is those involved in the polar case of unlimited leverage. Here the investor has perfect and costless access to loan markets and to facilities for recontracting his liabilities. An investor constrained by the condition $W = A - L$ can then gamble on his expectation that interest rates will rise by employing one or a combination of the following strategies: (1) recontract existing liabilities farther into the future; (2) create new long-term liabilities and hold the proceeds either as cash or in the form of short-term assets; and (3) sell long-term assets and hold the proceeds either as cash or in the form of short-term assets. Analogous strategies are easily stated for the case in which the investor wishes to gamble on his expectation that interest rates will fall. In either case, the point to be made is that these assumptions imply that A and L are *also* decision variables and that the investor can adjust the difference $Lm_L - Am_A$, subject only to the constraints that m_A and m_L be nonnegative and that $A - L = W$.

The opposite polar case is, of course, the trivial one where neither the sizes nor the time distributions of the asset and liability streams can be adjusted. The difference $Lm_L - Am_A$ is then a datum to the investor, and he can do nothing to alter the gamble on interest rate movements that any nonzero difference implies.

The assumptions of neither polar case hold in the real world. If he is willing to bear the costs, an investor can exchange short-term for long-term assets or the reverse. He can also recontract his liabilities, though neither costlessly nor perfectly. Within the narrow limits typically imposed by creditors and regulatory agencies, he can create liabilities to purchase assets.

Rather than attempt to deal with the variety of leverage constraints, recontracting arrangements, and patterns of transactions costs that are possible in any situation approximating the real world, I shall make the following simplifying assumptions in the model to be developed below: (1) both the size and the time distribution of the investor's liability stream are fixed, and (2) the investor can exchange assets costlessly in a continuum of maturities ranging from $t = 0$ (cash) to $t = T$.

If the investor in question is an institution such as a life insurance company, a bank, or an investment trust, the first assumption is satisfied in a world where a regulatory agency prohibits any recontracting of liabilities or leveraging operations involved, for example, in undertaking reinsurance contracts or in efforts to change the mix of deposits or accounts. The time distribution of any given volume of liabilities for such an institution would then be fixed. To be sure, the sizes of the asset and liability streams of the institution change in the normal course of its business as it sells new policies or acquires new accounts. But if we assume that the decision period is short enough so that no new sales nor additions to accounts are possible, then both the size and time distribution of liabilities are fixed as assumption (1) requires.

Taken together, assumptions (1) and (2) imply that A, L, and m_L are fixed for a given interest rate structure and that m_A can vary continuously on the closed interval $[0, T]$. I shall assume in addition that the fixed value of m_L, m_L^0, lies in the open interval $(0, T)$ so that the difference $Lm_L^0 - Am_A$ then lies in the interval

$$(Lm_L^0 - AT) \leqq (Lm_L^0 - Am_A) \leqq Lm_L^0. \qquad (9)$$

Return now to the definitions and assumptions of Section 2. Before proposing a decision rule and discussing its implications, I shall impose one further restriction on the problem. As stated, approximation (5) does not rule out the possibility of negative values of w. It is possible, therefore, that the investor can undertake a gamble on interest rate movements for which he can lose more than his current wealth, W. I shall arbitrarily rule out this possibility by assuming that the investor believes that the worst possible outcome of any decision he can make is one for which $w = 0$.[14] This assumption can be incorporated in the model, given the restrictions imposed earlier on A, L, m_A, and m_L by requiring that the random variable h lie in the closed interval:[15]

---

[14] This restriction is imposed to avoid complications in the use of the utility function and measure of risk aversion to be introduced below.

$$-(W/Lm_L{}^0) \leqq a \leqq h \leqq b \leqq W/(AT - Lm_L{}^0). \qquad (10)$$

But we should note that this restriction on h implies that w is strictly bounded as well. Statement (10) implies that

$$0 \leqq w \leqq \max[WAT/(AT - Lm_L{}^0),\ WAT/Lm_L{}^0] = w^*. \qquad (11)$$

Clearly, W belongs to this interval and reasonable assumptions about the relative sizes of the parameters assure that this restriction does not seriously limit the range of w.[16] For convenience, I shall refer to the set of all values of w satisfying these inequalities as the set D.

With these assumptions and restrictions on the problem we can now turn to a statement of the decision rule. Following Arrow[17] and others, I shall assume that the investor's tastes in wealth can be summarized by the von Neumann-Morgenstern utility function $u(w)$. The function u is unique up to an increasing linear transformation. In addition, I shall assume that u and its first two derivatives u' and u'' are continuous on the set D. If wealth is desirable, then $u' > 0$. If the investor is risk averse as well, then $u'' < 0$ everywhere on D.[18] These assumptions together imply that u is bounded and strictly concave on the compact set D.

The function

$$Z(w) = -u''(w)/u'(w) \qquad (12)$$

is a measure of the investor's risk aversion.[19] For a given value of w, $Z(w)$ is proportional to the premium over fair odds the investor will insist on to accept a bet of fixed size. This measure of risk aversion is invariant under linear transformations of the utility function chosen to represent the investor's tastes in wealth, and is continuous on the set D. The hypothesis that Z is decreasing on the set over which w is defined is an important source of empirical restrictions on the decision variables in models involving risk. It asserts that the decision maker's willingness to accept a bet of fixed size increases with his wealth in the sense that the size of the premium he demands over fair odds decreases.

The decision rule can now be stated as follows: Choose m_A to maximize the expected value of the utility of wealth. More formally, let M be the set of all values of m_A belonging to the interval

---

[15] Since $A > L > 0$, the assumptions leading to (9) imply that both $Lm_L{}^0$ and $(AT - Lm_L{}^0)$ are positive so that the ratios required in (10) exist.

[16] As the reader may verify, the maximum in (11) depends on the size of the ratio $AT/Lm_L{}^0$ (> 1). For values of this ratio of 2 or more, $w^* \geqslant 2W$.

[17] See [1].

[18] An investor is risk averse if he always refuses *any part* of a bet which is at best actuarially fair. He will, however, always accept *some part* of a bet which is actuarially favorable. See [1], pp. 92–93, 100.

[19] The function Z measures the investor's "absolute risk aversion." For a detailed discussion of the properties of Z and related measures, the reader should consult Arrow [1], pp. 93–96, and the references he cites.

$[0, T]$. Then the decision rule is that the investor choose that value of m_A in M that maximizes the function

$$F(m_A) = \int_a^b u[W + h(Lm_L^0 - Am_A)]f(h)dh. \qquad (13)$$

Since F is continuous and strictly concave on the closed and bounded set M, we are assured that F assumes an absolute maximum for some m_A in M. This maximum may occur at $m_A = 0$, $m_A = T$, or at some value of m_A in the interior of M. The conditions for these maxima are summarized in the following Proposition.[20]

Proposition 1: The function F has a maximum at a value $\overline{m}_A$ such that

(1) $\overline{m}_A = 0$ if and only if $F'(0) \leqq 0$
(2) $\overline{m}_A = T$ if and only if $F'(T) \geqq 0$
(3) $0 < \overline{m}_A < T$ if and only if $F'(\overline{m}_A) = 0$.

Statements (1) and (2) provide the conditions for corner solutions in the model. Although these conditions are interesting from a theoretical point of view,[21] I shall concentrate attention on interior solutions and, in particular, on the interior solution involving hedging or immunizing the balance sheet. Does risk aversion *always assure* hedging behavior on the part of the investor? If not, under what circumstances will he gamble on interest rate movements?

Statement (3) of the Proposition answers the first question. The investor hedges his balance sheet by chosing $m_A = Lm_L^0/A$ because then $w = W$. This value of m_A is optimal if and only if $F'(Lm_L^0/A) = -Au'(W)E(h) = 0$, i.e., if and only if $E(h) = 0$. Hence, the investor hedges *if and only if* he expects that interest rates are not going to change. Only a few essentially similar additional arguments are then needed to establish Proposition 2.

Proposition 2: Assume that F has a maximum at a value $\overline{m}_A$ in the interior of the set M. Then $Lm_L^0 - A\overline{m}_A \geqq 0$ (resp. < 0) if and only if $E(h) \geqq 0$ (resp. < 0).

Proposition 2 is a restatement of Redington's immunization rule. It replaces a rule stating that hedging is optimal under *any* circumstances a risk averse investor may face by a far more appealing one which directly relates his decision to hedge or to speculate to his expectations about interest rate movements. Under it, an investor who is risk averse immunizes his balance sheet when, and only when, he "expects" he has nothing to gain from speculating. In all other circumstances, immunization *is not* optimal investment behavior.

---

[20] Proofs of this Proposition and those to follow are sketched in the Appendix.

[21] Arrow [1], pp. 111 ff., shows that corner solutions are closely related to what he calls critical values of the investor's risk aversion. In the model Arrow discusses, an investor holds or does not hold some of his wealth in the form of a safe asset according as a certain measure of risk aversion exceeds or fails to exceed some critical value. Similar considerations in this model lead to conditions under which solutions occur at the corners. Since these arguments are quite complicated and add little to the main arguments, I shall omit them.

It is worth noting that this result is a simple consequence of risk aversion. The difference $x = Lm_L{}^0 - Am_A$ is the amount of his current wealth, W, the investor places at risk, i.e., x is the size of the bet he makes on interest rate movements. Choosing m_A, so that $x \neq 0$ when $E(h) = 0$ contradicts the assumption that the investor in question is risk averse, because then he is accepting a bet at actuarially fair odds. Similarly, choosing m_A so that $x = 0$ when $E(h) \neq 0$ means that the investor has refused *any part* of an actuarially favorable bet—a conclusion which also contradicts the assumption of risk aversion.[22]

The immunization rule contained in Proposition 2 links the investor's decision to speculate or to hedge to his subjective forecast of interest rate movements. As stated, however, it says nothing about the other factors which influence his decision. Certainly, the *extent* to which he is willing to speculate if $E(h) \neq 0$ depends on the magnitude of $E(h)$, i.e., x depends on $E(h)$. But the uncertainty with which he makes this forecast will also be an important element in his decision, as will his wealth (the size of the bet he is capable of making) and his aversion to risk. If we are willing to specify the functional form of $u(w)$, then Proposition 1 will enable us to solve for the optimal value of m_A as a function of the parameters and so allow us to take account jointly of all the factors governing the investor's decision. In order to avoid any peculiarities that may be introduced by the assumption of a particular utility function, I shall employ the more general comparative statics procedures in attempting to isolate the influence of each of these factors. Proposition 3 below provides the basis for an examination of the effect of changes in wealth on the investor's decision.[23]

Proposition 3: Assume that F has a maximum at a value $\overline{m}_A$ in the interior of the set M. Assume, further, that Z is decreasing on the set D. If we write $\overline{x} = Lm_L{}^0 - A\overline{m}_A$, then $\partial\overline{x}/\partial W \geqq 0$ (resp. < 0) if $\overline{x} \geqq 0$ (resp. < 0).

If the investor's utility function displays decreasing absolute risk aversion, i.e., if Z is decreasing on D, then Proposition 3 tells us that risky investment is not an "inferior good."

To illustrate, assume that $\overline{x} \neq 0$ initially. If the investor's subjective estimate of the odds on the gamble he was initially offered is independent of the level of his wealth, then any (small) change in W changes Z and, hence, the optimal value of x: the premium over fair odds the investor originally insisted on to accept the bet $\overline{x}$ changes, and $\overline{x}$ is no longer optimal. If, for example, $\overline{x}$ is positive initially and W increases (decreases), then Z decreases (increases) and with it the premium over fair odds the investor requires to accept a bet of size $\overline{x}$. He responds by increasing (decreasing) x through the appropriate change in m_A.

---

[22] Cf. note 14, above.

[23] See the Appendix for the details of the way in which wealth is assumed to change in the model.

The case in which $\bar{x}$ is negative initially is easily interpreted along similar lines, but the hedging case may require a comment. If $\bar{x}$ is initially zero, then Proposition 2 states that the investor's subjective estimate of the odds on any bet he could have made on interest rate movements was at best only actuarially fair odds. Any change in wealth that does not change these odds changes only the size of the bet he is capable of making. The investor clearly has no reason to change $\bar{x}$ as Proposition 3 formally asserts.

Isolating the influence of expectations on the investor's decisions is a more complicated problem. A change in the investor's forecast, $E(h)$, can be represented by a shift in his subjective probability density, $f(h)$. But a shift in $f(h)$ that increases $E(h)$ can be one that changes its other features as well. If, as seems natural, we take the variance of h as a measure of the uncertainty an investor attaches to his forecast, then a shift in $f(h)$ representing an investor's revision in $E(h)$ can be accompanied by an increased, a decreased, or a constant "degree of uncertainty." Since the investor's response to each of the possible mixtures of a given revision in $E(h)$ with each of the possible changes in uncertainty is likely to be very different, some classification scheme for "changes in expectations" is clearly required.

Propositions 4 and 5 below are based on the following two types of changes in expectations: (1) a shift in $f(h)$ which changes the mean of h but preserves its variance, and (2) a shift in $f(h)$ which changes the variance of h but preserves its mean. The first of these changes can be regarded as a pure revision in the investor's forecast: the second as a pure revision in the degree of his uncertainty. Leaving mathematical details to the Appendix, I shall now state and interpret Proposition 4.

Proposition 4: Assume that F has a maximum at a value $\bar{m}_A$ in the interior of the set M, and write $\bar{x} = Lm_L{}^0 - A\bar{m}_A$. Let θ be a parameter defined such that an increase (a decrease) in θ increases (decreases) the mean of the random variable h but preserves its variance. Then the change in $\bar{x}$ induced by a small change in θ can be written as the sum

$$\left.(\partial\bar{x}/\partial\theta)\right|_{\substack{\text{wealth}\\ \text{constant}}} = \left.(\partial\bar{x}/\partial\theta)\right|_{\substack{\text{expected}\\ \text{utility}\\ \text{constant}}} + \bar{x}\left.(\partial\bar{x}/\partial W)\right|_{\substack{\theta\\ \text{constant.}}}$$

The first term of this sum is positive. If Z is decreasing on the set D, then the sum is positive.

Suppose that for some reason an investor revises his forecast of interest rate movements. If the degree of his uncertainty remains the same, then Proposition 4 states that the accompanying change in $\bar{x}$ can be treated as an analogue to a price change in the theory of consumers' choice and can be similarly decomposed. The disguised change in wealth a pure revision in his forecast entails can be netted out by a compensating variation in the investor's wealth just sufficient to hold his expected utility constant at its initial equilibrium value. The resulting term, which is always positive in sign, can be regarded as the substitution effect of a pure revision in the investor's forecast. When he is initially hedging, the wealth effect of this revision is zero, and the sign of the induced change in $\bar{x}$ is unambiguously positive. If the investor is decreasingly risk averse as his wealth increases, then Proposition 3 assures us that any wealth effect due to a pure revision in his forecast will always be nonnegative. We can then conclude that under decreasing absolute risk aversion, the size, x, of the bet the investor is willing to make on interest rate movements is an increasing function of $E(h)$ in a neighborhood of its initial equilibrium value, $\bar{x}$.

A Slutsky decomposition into substitution and wealth effects is also possible in the case of a pure revision in the degree of uncertainty. Assume that the investor revises the degree of uncertainty with which he makes his forecast. If $E(h)$ remains the same, then the resulting change in $\bar{x}$ can be represented as the sum of two terms. By wealth compensating in such a way as to hold expected utility constant at its initial equilibrium value, the substitution effect of a pure revision in the degree of uncertainty can be isolated. This term is zero when the investor is hedging, and has the sign of $-\bar{x}$ otherwise. Hence, a wealth compensated upward (downward) revision in the degree of uncertainty with which the investor makes his forecast is accompanied by a decrease (increase) in the size of the bet he is willing to make on interest rate movements. If the investor is decreasingly risk averse as his wealth increases, then the wealth effect of a pure revision in the degree of his uncertainty reinforces the substitution effect. These results are summarized in Proposition 5.

Proposition 5: Assume that F has a maximum at a value $\bar{m}_A$ in the interior of the set M, and write $\bar{x} = Lm_L{}^0 - A\bar{m}_A$. Let θ be a parameter defined such that an increase (a decrease) in θ increases (decreases) the variance of the random variable h but preserves its mean. Then the change in $\bar{x}$ induced by a small change in θ can be written as the sum

$$(\partial\bar{x}/\partial\theta)\Bigg|_{\substack{\text{wealth}\\ \text{constant}}} = (\partial\bar{x}/\partial\theta)\Bigg|_{\substack{\text{expected}\\ \text{utility}\\ \text{constant}}} + E(h)\bar{x}(\partial\bar{x}/\partial W)\Bigg|_{\substack{\theta\\ \text{constant.}}}$$

The first term of the sum has the sign of $-\bar{x}$. The second term of the sum has the sign of $E(h)$ if Z is decreasing on the set D.

4. Concluding notes

■ In conclusion, I would like to make two comments on the model developed in Sections 2 and 3.

The first concerns empirical tests of the model. As stated, the model links the investor's choice of m_A to his expectations, to his wealth, and to certain characteristics of his utility function for wealth. Any empirical specification of the model must, of course, deal with the measurement problems these features of the model present. But surprisingly, the major problem is not obtaining an empirical specification as such, but rather obtaining data on the duration m_A.[24] Calculating the duration of a 30-year coupon bond is not an easy task. Calculating the duration for even a very simple bond portfolio (much less a statistically meaningful sample of portfolios) is a major undertaking.

To my knowledge, there are no data available on investment portfolios reporting values of m_A. The average maturity of a bond portfolio is, to be sure, a proxy variable for m_A, but it is very likely to be one which is seriously biased upward if the bonds in the portfolio have large coupons and/or long periods to maturity.[25] Even if sufficient data were available on a cross section sample of bond portfolios, one could not be sure that differences in m_A are not disguised by coupon differences in bonds held by different members of the cross section. A careful test of the model would, in short, involve a major research effort.

The second comment concerns the relevance of the model to studies of interest rate structures. Hicks[26] maintained that the demand side of the "forward market" for loans can be expected to display a constitutional weakness, that "most people (and institutions) would prefer to lend short" and so must be tempted to lend long by an interest rate sufficiently large to compensate for the extra risk incurred. Variations on this argument have long been debated by those who see loan markets as being dominated by speculators indifferent to maturities of loans and by those who argue that loan markets are segmented by the demands of those investors who prefer to lend short and by those who prefer to lend long. The model implies that Hicks' argument is only valid under the assumption that lenders have, on average, short liabilities so that $m_L{}^0$ is small. Indeed, in a world in which loan markets are dominated, for example, by life insurance companies, the reverse would be true: lenders would have to be tempted into lending *short* to compensate for the extra risk incurred. The model supports a form of the so-called segmentation hypothesis, but one based on the nature of the lenders' liabilities and their expectations and risk aversion rather than on any *natural* preference for some lenders to lend short and others to lend long.[27]

Appendix

■ Proofs of the propositions contained in the text are outlined in the paragraphs below numbered accordingly. For reasons of space, I have omitted proofs easily obtainable elsewhere. References are supplied in these instances.

---

[24] Several empirical specifications based on the equilibrium condition (3) of Proposition 1 in the text are possible, though all require data on m_A. I have experimented with one of them (with average maturity as a proxy for m_A) using some data on life insurance portfolios acquired by the Temporary National Economic Committee in 1940. The results, though encouraging, are too thin (and dated) to warrant the space required for a formal presentation. I will supply them to any interested reader.

[25] The extent of this bias is illustrated in Fisher and Weil [2].

[26] In [5], p. 146.

[27] It is worth noting here that an investor can attempt to reduce risk by an effort to change $m_L{}^0$. For example, an investor for whom $m_L{}^0$ is small could reduce the risk of lending long by inducing his creditors to recontract his liabilities. The attractive rates offered on time certificates of deposit by banks in recent years may be a case in point.

(1) All three statements of Proposition 1 follow easily from the strict concavity of the function F.

(2) These are essentially the same arguments as those of Arrow.[28]

(3) (a) I shall assume that the *only* changes in W that can occur exogenously are those due to "parallel displacements" of $A(t)$ and/or $L(t)$, i.e., $A^*(t) = \lambda A(t)$ and/or $L^*(t) = \lambda L(t)$ and $\lambda > 0$. The durations m_A and m_L defined in (3) in the text are then unchanged.

(b) Arrow's arguments[29] can be employed to show that when $x = Lm_L{}^0 - Am_A \geqq 0$, decreasing absolute risk aversion implies that $E[hu''(w)] \geqq 0$. A simple extension of his proof for the case when $x < 0$ is admissible can be made to show that $\bar{x} \cdot E[hu''(w)] \geqq 0$.

(c) Differentiate $E[hu'(W + h\bar{x})] = 0$ partially with respect to W (keeping in mind that $\bar{x}$ is a function of W in a sufficiently small neighborhood of the initial value of W). Solving,

$$\partial\bar{x}/\partial W = -E[hu''(\bar{w})]/E[h^2u''(\bar{w})]. \qquad \text{(A1)}$$

The sign restrictions in the Proposition then follow from paragraph (b) above.

(4) (a) The transformation of variable $h(\theta) = h + \theta$ will yield the required variance preserving shift.

(b) Differentiate $E\{u[W + h(\theta)\bar{x}]\}$ partially with respect to θ and assume that W is also a function of θ. The required compensation in W necessary to hold $E[u(\bar{w})]$ constant is $\partial W/\partial\theta = -\bar{x}$.

(c) Differentiate $E\{h(\theta)u''[W + h(\theta)\bar{x}]\} = 0$ partially with respect to θ and wealth compensate. Evaluating derivatives at $\theta = 0$,

$$\partial\bar{x}/\partial\theta = -E[u'(\bar{w})]/E[h^2u''(\bar{w})] > 0. \qquad \text{(A2)}$$

The sign restriction follows from the assumptions about the derivatives of u.

(d) Differentiate the expression in (c) above partially with respect to θ, but now hold W constant. Evaluating derivatives at $\theta = 0$ and using (A1) and (A2), we obtain the expression in the text.

(5) (a) The transformation of variable $h(\theta) = E(h) + (1 + \theta) \cdot [h - E(h)]$ will yield the required mean preserving shift.

(b) The wealth compensation necessary to hold expected utility constant at its initial equilibrium level is $\partial W/\partial\theta = \bar{x}E(h)$.

(c) Differentiate $E\{h(\theta)u'[W + h(\theta)\bar{x}]\} = 0$ partially with respect to θ and wealth compensate. Evaluating derivatives at $\theta = 0$,

---

[28] In [1], p. 99.

[29] In [1], p. 119.

$$\partial\bar{x}/\partial\theta = E(h)E[u'(\bar{w})]/E[h^2u''(\bar{w})] - \bar{x}. \tag{A3}$$

The sign restrictions follow from Proposition 2.

(d) Differentiate the expression in (c) above partially with respect to θ, but now hold W constant. Evaluating derivatives at $\theta = 0$ and using (A1) and (A2), we obtain the expression in the text.

References

1. Arrow, K. J. *Essays on the Theory of Risk Bearing.* Chicago: Markham Publishing Company, 1971.
2. Fisher, L. and Weil, R. L. "Coping with the Risk of Interest Rate Fluctuations: Returns to Bondholders from Naive and Optimal Strategies." *Journal of Business,* Vol. 44, No. 4 (October 1971), pp. 111–118.
3. Grove, M. A. "A Model of the Maturity Profile of the Balance Sheet." *Metroeconomica,* Vol. 18, No. 1 (April 1966), pp. 40–55.
4. Hickman, W. B. "The Term Structure of Interest Rates." Unpublished manuscript, 1942.
5. Hicks, J. R. *Value and Capital.* Oxford: Clarendon Press, 1946.
6. Hopewell, M. H. and Kaufman, G. G. "Bond Price Volatility and Term to Maturity: A Generalized Respecification." *The American Economic Review,* forthcoming.
7. Macaulay, F. R. *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the U. S. since 1856.* New York: National Bureau of Economic Research, 1938.
8. Redington, F. M. "Review of the Principles of Life Office Valuations." *Journal of the Institute of Actuaries,* Vol. 78, No. 3 (1952), pp. 286–340.
9. Samuelson, P. A. "The Effect of Interest Rate Increases on the Banking System." *The American Economic Review,* Vol. 35, No. 1 (March 1945), pp. 16–27.
10. Wallas, G. E. "Immunization." *Journal of the Institute of Actuaries Students' Society,* Vol. 15 (1959), pp. 345–57.
11. Weil, R. L. "Macaulay's Duration: An Appreciation." *Journal of Business,* Vol. 46, No. 4 (October 1973), pp. 589–592.

The behavior of a firm subject to stochastic regulatory review: comment

William T. Ziemba

Associate Professor of Management Science
Faculty of Commerce and Business Administration
University of British Columbia

In his analysis of the behavior of a firm subject to stochastic regulatory review in an earlier issue of this Journal, Klevorick uses some "well-known results" concerning the strict concavity of composite functions to prove the unique existence of the firm's optimal research and investment levels in each period. Unfortunately some of these "well-known results" are not generally correct. Hence the proof of Theorem 2 is incorrect as stated. See Klevorick's accompanying reply for his corrections of the proof.[1]

1. Counterexamples and salvaging the strict concavity results

■ It develops that Klevorick's Lemmas 1 and 2 are incorrect, while the remark and Lemma 3 are correct.[2] Lemma 2 reads as follows:

Lemma 2: Let x_1, x_2, x_3, x_4, x_5 be real-valued variables and let $u(x_1, x_2, x_3)$ and $w(x_4, x_5)$ be real-valued functions. If (i) $u(x_1, x_2, x_3)$ is strictly concave, (ii) $u(x_1, x_2, x_3)$ is everywhere nonincreasing in x_1 for x_2, x_3 fixed, and (iii) $w(x_4, x_5)$ is convex, then $u[w(x_4, x_5), x_2, x_3]$ is a strictly concave function of x_2, x_3, x_4, and x_5.

As a counterexample[3] to Lemma 2 let $u(x_1, x_2, x_3) = -x_1^2 - x_2^2 - x_3^2$ and $w(x_4, x_5) = x_4 + x_5$, where each $x_i \geqq 0$. Hence $u[w(x_4, x_5), x_2, x_3] = -(x_4 + x_5)^2 - x_2^2 - x_3^2 \equiv \psi(x)$, where $x \equiv (x_2, x_3, x_4, x_5)$. Now if $x^1 = (0, 0, 1/2, 1/2)$, $x^2 = (0, 0, 1/4, 3/4)$, and $x^3 = (1/2)x^1 + (1/2)x^2$, it follows that $\psi(x^3) = 1/2\psi(x^1) + 1/2\psi(x^2)$, and that ψ, therefore, is not strictly concave. The essence of this example is that $w[\lambda x_4^1 + (1 - \lambda)x_4^2, \lambda x_5^1 + (1 - \lambda)x_5^2]$ is a constant for all $\lambda \in [0, 1]$. Hence one can construct similar counterexamples for numerous alternative u.

Lemma 1 reads as follows:

Lemma 1: Given the maximization problem

$$\underset{x}{\text{Maximize}}\,[f(x, y)\,|\,g(x) \geqq b],$$

Thanks are due to the National Research Council for financial support and to A. Amershi, P. MacAvoy, and C. Swoveland for helpful comments on an earlier draft of this note.

[1] In [3].

[2] See [2], p. 84.

[3] The error in assuming the validity of results like Lemma 2 is common, see, e.g., Fama [1], footnote 18.

where x is an n-vector, y is an r-vector, $f(x, y)$ is a strictly concave real-valued function, $g(x)$ is an m-vector-valued concave function, and b is an m-vector, the function $\psi(b, y)$ defined by

$$\psi(b, y) \equiv \operatorname*{Max}_{x \cdot \ni \cdot g(x) \geqq b} f(x, y)$$

is a strictly concave function of b and y.

Lemma 1 may be invalidated by letting $n = r = m = 1$, $f(x, y) = x_1 - 1/2x_1^2 - y_1^2$, and $g(x) = x_1 \geqslant b$. Let $(b^1, y^1) = (0, 0)$, $(b^2, y^2) = (1, 0)$, and $(b^3, y^3) = \lambda(b^1, y^1) + (1 - \lambda)(b^2, y^2)$. Then $\psi(b^3, y^3) = \lambda\psi(b^1, y^1) + (1 - \lambda)\psi(b^2, y^2) = 1/2$ for all $\lambda \epsilon [0, 1]$. Hence ψ is not strictly concave. Such counterexamples arise when the optimal x vector is independent of b for fixed y. This difficulty is not crucial in the problem posed in the remark. Indeed, the validity of the remark may be verified by using an argument analogous to that in the proof of Lemma 1′. Lemma 1 may be salvaged with the additional stringent assumption that $x^*(b)$, the optimal solution for given b, is one-to-one, that is, there is a unique b for every given value of $y[=x^*(b)]$. There do not seem to be general sufficient conditions to guarantee that this stringent assumption is satisfied.

Lemma 1′: Let $\psi(b, y) \equiv [\max_{x} f(x, y) | g(x) \geqq b]$.

Suppose x is a real n-vector, y is a real r-vector, $f(x, y)$ is a strictly concave real-valued function, $g(x)$ is an m-vector-valued concave function, b is a real m-vector and $x^*(b)$ exists and is one-to-one. Then ψ is strictly concave.[4]

Lemma 2 may be salvaged if one makes the stringent assumption that the function $w(x_4, x_5)$ is one-to-one, that is there is a unique pair (x_4, x_5) for every given value of $y = w(x_4, x_5)$. Alternatively the result obtains if w is strictly convex. The following Lemma presents what seem to be the weakest assumptions that will guarantee the strict concavity of the composite function.

Lemma A: Let $y \equiv (y_1, \ldots, y_m)$ be real-valued variables, and $x_i \equiv g_i(y_1, \ldots, y_m)$, $i = 1, \ldots, n$ and $u(x_1, \ldots, x_n)$ be real-valued functions. Suppose at least one of the following set of assumptions is satisfied:

(1) u is strictly concave and nondecreasing, each g_i is concave, and some g_i is one-to-one;

(2) each g_i is concave and some g_i, say g_k, is strictly concave, u is concave, nondecreasing, and strictly increasing in g_k. Then $f(y) \equiv u[g_1(y), \ldots, g_n(y)]$ is strictly concave.

The assumption in (2) is equivalent to u's being strictly concave in g_k for fixed g_i, $i \neq k$. Lemma A has many variants.[5] For example, (1) may be changed to (1′): u is strictly concave and nonincreasing, each g_i is convex, and some g_i is one-to-one. The validity of

---

[4] Due to space limitations, the proofs of Lemmas 1′ and A could not be included here; they appear in [5].

[5] See Mangasarian [4] for ideas to further extensions based on his pioneering treatment of various (nonstrict) composite results involving concave and related functions.

Lemma 3 may now be gleaned from the fact that x_2 and x_3 are (trivial) one-to-one functions of x_2 and x_3, respectively. Similarly (2) may be changed to (2′): each g_i is convex and some g_i, e.g., g_k, is strictly convex, u is concave, nonincreasingly, and strictly decreasing in g_k.

References

1. Fama, E. F. "Multiperiod Consumption-Investment Decisions." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 163–174.
2. Klevorick, A. K. "The Behavior of a Firm Subject to Stochastic Regulatory Review." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 57–88.
3. ———. "The Behavior of a Firm Subject to Stochastic Regulatory Review: Correction." *The Bell Journal of Economics and Management Science*, Vol. 5, No. 2 (Autumn 1974), pp. 713–714.
4. Mangasarian, O. L. "Convexity, Pseudo-Convexity, and Quasi-Convexity of Composite Functions." *Cahiers du Centre d'Etudes de Recherche Operationelle*, Vol. 12 (1970), pp. 114–122.
5. Ziemba, W. T. "Note on 'The Behavior of a Firm Subject to Stochastic Regulatory Review.'" Working Paper No. 231, University of British Columbia, Faculty of Commerce, Vancouver, 1974.

The behavior of a firm subject to stochastic regulatory review: correction

Alvin K. Klevorick

Associate Professor of Law and Economics
Yale University

In the following note the author supplies corrections for Lemmas 1 and 2 which appeared in his earlier article in this Journal. The revised Lemmas can be used to correct his original proof of Theorem 2.

■ W. T. Ziemba's counterexamples[1] are clearly correct and to the point. The results cited as Lemmas 1 and 2 in my original paper are incorrect, while the Remark and Lemma 3 are correct. As a result, the proof of Theorem 2 presented in the original paper is incorrect. Theorem 2 itself, however, *is* correct as it stands, as is the basic strategy of the proof. The major changes required to correct the proof follow.

☐ Substitute the following for Lemmas 1 and 2, and the Remark:

Lemma 1. Given the maximization problem

$$\underset{x}{\text{Maximize}}\ f(x, y) \qquad \text{Subject to: } g(x) \geqq b,$$

where x is an n-vector, y is an r-vector, $f(x, y)$ is a concave real-valued function, $g(x)$ is an m-vector-valued concave function, and b is an m-vector, the function $\psi(b, y)$ defined by

$$\psi(b, y) \equiv \underset{x \ni g(x) \geqq b}{\text{Max}} f(x, y)$$

is a concave function of b and y.

Remark. For the special case in which $y = b$ (so that $r = m$), the result of Lemma 1 is that the function $\psi(b)$ is concave where

$$\psi(b) \equiv \underset{x \ni g(x) \geqq b}{\text{Max}} f(x, b)$$

Lemma 2. Let x_1, x_2, x_3, x_4, x_5 be real-valued variables and let $u(x_1, x_2, x_3)$ and $w(x_4, x_5)$ be real-valued functions. If (i) $u(x_1, x_2, x_3)$ is concave, (ii) $u(x_1, x_2, x_3)$ is everywhere nonincreasing in x_1 for x_2, x_3 fixed, and (iii) $w(x_4, x_5)$ is convex, then $u[w(x_4, x_5), x_2, x_3]$ is a concave function of x_2, x_3, x_4, and x_5.

☐ Instead of proving the four statements at the top of p. 85[2] for each j, the following four propositions are proven for each j:

(a) $H_j(q, K, A)$ is strictly concave in q, K, A.
(b) $V_j(q, K, A)$ is concave in q, K, A.
(c) $V_j(q, K, A)$ is everywhere nonincreasing in q for K, A fixed.
(d) $V_j[\phi(K, A), K, A]$ is concave in K and A.

---

[1] In [2].
[2] In [1].

Essentially, in using Lemmas 1 and 2 to prove Theorem 2 in the original paper, I used incorrect statements that were stronger than what I actually needed. The changes given here necessitate a number of other alterations in the proof of Theorem 2—almost all of them involving appropriate deletions of the word "strictly." Space limitations preclude indicating all of these changes here, but I shall be happy to provide a complete copy of the corrected proof on request.

References

1. KLEVORICK, A. K. "The Behavior of a Firm Subject to Stochastic Regulatory Review." *The Bell Journal of Economics and Management Science,* Vol. 4, No. 1 (Spring 1973), pp. 57–88.
2. ZIEMBA, W. T. "The Behavior of a Firm Subject to Stochastic Regulatory Review: Comment." *The Bell Journal of Economics and Management Science,* Vol. 5, No. 2 (Autumn 1974), pp. 710–712.

Lieb's *Freight Transportation: A Study of Federal Intermodal Ownership Policy*

Reviewed by
Harold A. Horowitz

Graduate Student in Law
Columbia University

■ Although fully integrated transportation companies function elsewhere in the world, one firm's ownership of all or some of the major modes of transportation has been excluded by law from the structure of American enterprise. Robert Lieb traces the sources of legal and economic opposition to intermodal transportation in his timely and important book.[1] Given the fact that American legal thinking embodies the notion that regulation of the transportation industry is beneficial to the public interest,[2] the corporate structure of this industry should be considered separately from questions of safety standards and pricing. Clearly, the regulatory attitude must be geared to promoting the most efficient manner of organization in the transportation industry. The important issue which Lieb's work raises is whether it is wise to perpetuate the current legal position of the transportation industry with regard to intermodal ownership: do modern economic theory and empirical information justify the strict regulatory policies which constrain the formation of integrated transportation companies?

Lieb states that the purpose of his work is to provide "a comprehensive legal and economic analysis of intermodal ownership which might be of value in determining future federal policy concerning this important issue."[3] His clearly detailed history of federal transportation policy contributes to this purpose.

The primary reason for the opposition of Congress and the regulatory agencies to intermodal ownership by a single transportation company was the contention that, given the relative size and scope of rail services, transportation companies would promote railroad domination and would weaken the competitive position of other modes of transportation. Although today the various possible combinations include air-motor, motor-rail, motor-water, and water-rail, the most extensive attempts have traditionally involved rail-motor combinations.

In 1912 the Panama Canal Act was passed to impede rail-controlled water operations or, more specifically, to impede combinations between railroads and barge companies. The conceptual basis for this action was stated in the Report of the House Committee on Interstate Commerce which accompanied the Bill:

---

[1] [5].

[2] See Lewis [4], Pegrum [7], and U. S. Senate Committee on Interstate and Foreign Commerce [11]. See Mertins [6] for a more extensive coverage of the issues addressed in this review.

[3] [5], p. 11.

> The proper function of a railroad corporation is to operate trains on its tracks, not to occupy the waters with ships in mock competition with itself, which in reality operate to the extinction of all genuine competition.[4]

The concern which prompted passage of this Bill was that railroad companies, with their great financial strength, could bar the entry of independent water carriers on railroad-controlled barge routes. The House Report emphasized that the Panama Canal Act was passed to exclude all railroad-owned shipping from competition with their rail service in coastwise trade, to outlaw the connection between railroads and water carriers in through routes, and to prohibit setting joint rates. Thus, the thrust of Congressional thinking was that the activities of the different modes of transportation should be mutually exclusive.

The Interstate Commerce Commission implemented the provisions of the Panama Canal Act in the Lake Lines Application case of 1915.[5] The Commission denied the major railroads permission to continue operation of steam vessels on the Great Lakes. The decision was made because the Commission felt that the boats were not being operated primarily to develop water transportation and that the water rates were too high to permit potential competition between the two forms of transport. These charges were accurate inasmuch as the more powerful railroads could drive out new competitors by lowering rates or increasing service in a manner not feasible for a new company with little capital.

More recently, the Commission denied the joint application of the Southern Pacific Company and the Illinois Central Railroad to acquire control of the John Hay Company which was a major barge operator on lines parallel to those of the railroads.[6] The Commission held that the barge line would be managed to serve the interests of the railroad and would subsequently direct traffic to all-rail movements and argued that the proposed combination would make the Hay Company so powerful that the competition the Commission had to preserve among the other water carriers would be precluded. In contrast to its historic position, the Commission has taken a different stance by permitting the Southern Company to engage in a limited contract carriage of coal by barge on the Ohio and Tennessee river systems.[7]

In other areas of joint ownership, regulatory and statutory actions have severely limited the scope of such combinations. Intermodal ownership was limited by the Motor Carrier Act of 1935, and was incorporated into Part Two (II) of the Interstate Commerce Act. The Interstate Commerce Commission has tended to protect existing motor operations, although it has allowed a limited amount of rail-motor combinations. However, under the Transportation Act of 1940, only railroads are restricted in their ability to control trucking operations. In terms of intermodal combinations between air carriers and surface carriers, it is interesting to note that although two different regulatory agencies—the CAB and the ICC—and different statutory provisions are involved, there is no major shift in attitude with regard to the formation of transportation companies. The Civil Aeronautics Board in its effort to promote intermodal competition has pursued policies to block integration of transportation companies. The Board has maintained modal segregation in its effort to preserve air-rail competition. The policies of the CAB have been guided by the second provision of Section 408 (b) of the Federal Aviation Act passed in 1938. Since then, the Board has never permitted a combination between air transport companies and surface carriers. It has not once found that surface carriers could use aircraft to the public advantage as part of its operations.

---

[4] U. S. Congress, "Report from Committee on Interstate and Foreign Commerce," H. R. 423, 62nd Congress, 2nd Session, 1912, pp. 12–13.

[5] *Lake Lines Applications under the Panama Canal Act*, 33 I.C.C. 699 (1915).

[6] Finance Docket No. 20940, *Illinois Central Railroad Control—John I. Hay Co.*, 317 I.C.C. 39 (1962).

[7] See Finance Docket No. 26310, *Application of Southern Railway Company under Section 5 (15) of the Interstate Commerce Act*, I.C.C., August 17, 1970.

The area where combinations have been permitted, without much difficulty, is in air forwarding ownership. A number of surface carriers have gained control of air-freight forwarding activities. This particular form of combination has been permitted to stimulate the growth of the air-freight industry by infusing outside financial resources and by easing access to a wider range of delivery points. The significance of this action is that there is a regulatory cognizance of the need to promote intermodal coordination in cases when it could lead to growth in the transportation industry. The challenge facing transportation economists and planners today is to determine whether there are real benefits to be derived from encouraging corporate combinations involving different modes on a much larger scale.

In summary, there have been extensive legislative restrictions created to govern intermodal ownership. However, most of these laws were enacted during the period when railroads were the dominant force in American freight transportation. Moreover, the financial strength of railroad companies was relatively so great that they might have been able to control the activities of any newly formed transportation companies that would emerge, if permitted. Thus, it is noteworthy that most of the legislation was directed at preventing railroads from taking over other modes of transportation. There was no comprehensive legislative overview with regard to the broader question of transportation coordination and other forms of transportation combinations.[8]

From this context and from the historical perspective, it is appropriate to consider the theoretical debate as to the merits of forming transportation companies in America. Although Lieb enumerates and assesses the major arguments of each position, he does not demonstrate their respective merits. The arguments in favor of the integrated transportation company center around the potential efficiencies to be derived from intermodal combination. Greater efficiency could result from the administrative savings of one-stop transportation services, i.e., matching the most appropriate mode with a particular transport need. Another potential benefit is the ability to channel capital into moving freight more speedily, rather than into advertising one particular mode of transportation. Moreover, transfer costs of changing modes of transport could be reduced and idle freight could be curtailed by coordinating the entire transportation system. Thus the most efficient transportation network would be made available to all potential customers. The key arguments against further integration of the transportation modes are that such organization would curtail the development of new modes, that it would channel the flow of goods to rail operations and therefore create more monopolistic conditions, and that management would find it difficult to operate all modes in a common interest.[9] However, the fact that railroads are no longer financially dominant in the transportation industry destroys the argument that they would dominate the transportation companies that might be formed. Although Lieb indicates the weaknesses of the arguments against multimodal ownership, he fails to present a strong case for the efficiency gains which could be derived from piggybacking, or integrating certain complementary transportation services. Moreover, though much has been written on the economics of piggybacking,[10] there have been only sparse theoretical insights into the benefits from an organizational and marketing point of view.[11]

---

[8] There is currently more attention being focused upon the regulatory environment which surrounds the transportation industry. The most notable example of such an effort by Congress is "The Transportation Regulatory Modernization Act of 1971," H.R. 11826; S. 2842. This bill recommends a move toward less regulation and more dependence on competitive forces.

One of the most appropriate means of assessing the potential value of the integrated transportation company is to consider the performance of such companies now in existence. At the present time there are two intermodal transportation companies in Canada —Canadian National and Canada Pacific. Without any qualitative and quantitative analyses, which are beyond the confines of a review, one may look at the effects of these companies on the transportation system by considering the Canadian experience with intermodal ownership. A Canadian Pacific executive, Harvey Romoff, asserted:

> Certainly the Canadian practitioners of multimodal ownership are enthusiastic supporters of the concept. I would say that intermodal transportation is the wave of the future. It has tremendous benefits to both transportation users and suppliers and hence to the economy overall.[12]

He indicated that multimodal ownership helps the user by enabling him to deal with a pool of transportation services, which results in cost benefits and a more efficient physical distribution system. Although Romoff's address did not offer conclusive statistical evidence for his assertions and did not examine the multimodal transportation company within a theoretical micro framework, Romoff's strong support of Canadian Pacific's organization is of interest. Clearly, however, assessing the potential benefits of an integrated transportation company merits further research.[13]

---

[9] For some insight into the opposing arguments, see Pegrum [7] and Beardsley [1].

[10] See Friedlaender [3] and Merrill J. Roberts and Associates [8].

[11] See Farris and McElhiney [2], pp. 356–361, and Wilson [12].

[12] See Romoff [9].

To improve our conceptual understanding of this problem tools must be developed to evaluate theoretically organizations that provide integrated services and to measure empirically their performance. The line of analysis should consider the potential economics of joint production functions, the extent to which more than one mode is used to transport each shipment between designated points, and the precise breakdown among the different modes of transportation per shipment of goods. Beyond this, vigorous empirical analysis of international experiences with these integrated companies should be conducted. Unfortunately, there is currently no data base adequate for such a study because of the forced separation of the modes even in a given organization.

As a guide to policy, the key question which emerges rather clearly is whether the different modes of transportation are substitutes for each other or are complementary services. If they are complementary services, the extent to which they complement each other must be determined. The traditional attitude has been that the various modes of transportation services are substitutes and that they should be segregated by law to maintain the viability of each particular form of service. From the regulatory angle, there is a three-fold problem which must be considered. The most fundamental question is whether regulation is beneficial to the economy with regard to any sector. The next question assumes a positive posture with regard to the first issue, and yet, queries whether the transportation industry should be regulated with respect to combinations and mergers, even within an active regulatory environment. Finally, although the problems of mergers and consolidations in general have not been thoroughly analyzed, the antitrust laws—particularly Section 7 of the Clayton Act—are designed to promote free competition by preventing anticompetitive mergers and combinations in the American economy. It seems to this writer that since the goal of public policy is primarily to promote free economic activity, which by definition presumes a relative ease of entry into an industry, there is no *a priori* reason to assume that transportation mergers could not be handled within the antitrust framework. The advantage of such an approach is that no *ipso facto* judgments are made concerning the desirability of such combinations, and yet anticompetitive combinations would be prohibited and enforced through the existing methods.

Lieb's stimulating and interesting work did not competely fulfill his aim of providing a comprehensive economic analysis of the concept of intermodal ownership. Yet, he has made an invaluable contribution to the formulation of future policy by analyzing the determinants of traditional policy concerning intermodal ownership. This work offers the background as to the policies which have been in force and the main reasons for these policies. It is now left to legislators, lawyers, and economists to evaluate whether these policies adequately meet current transportation needs.

---

[13] The most suggestive and interesting effort thus far has been Suelflow and Hille [10].

References

1. Beardsley, P. "Intermodal Ownership of Transportation Facilities—A Motor Carrier View." Address delivered to the Federal Bar Association, Washington, D. C., March 20, 1969.
2. Farris, M. and McElhiney, eds. *Mode[illegible]ransportation: Selected Readings.* Boston: Houghton-Mifflin, 1967.
3. Friedlaender, A. *The Dilemma of Freight Transport Regulation.* Washington, D.C.: The Brookings Institution, 1969.
4. Lewis, B. "Ambivalence in Public Policy toward Regulated Industries." *The American Economic Review,* Vol. 53, No. 1 (March 1963), p. 38.
5. Lieb, R. *Freight Transportation: A Study of Federal Intermodal Ownership Policy.* New York: Praeger Publishers, 1972.
6. Mertins, H. *National Transportation Policy in Transition.* Lexington, Mass.: D. C. Heath, 1972.
7. Pegrum, D. *Transportation: Economics and Public Policy.* Homewood, Ill.: Richard D. Irwin, 1963.
8. Roberts, Merrill J. and Associates. *Intermodal Freight Transportation Coordination: Problems and Potential.* Washington, D. C.: U. S. Government Printing Office, 1967.
9. Romoff, H. "Multimodal Transportation: Some Comments Based on the Canadian Experience." Address before the Transportation Research Forum, New Orleans, October 22–24, 1970.
10. Suelflow, J. and Hille, S. "The Transportation Company: An Economic Argument for Intermodal Ownership." *Land Economics,* Vol. 46, No. 3 (August 1970), pp. 275–286.
11. U. S. Senate Committee on Interstate and Foreign Commerce, Special Study Group on Transportation Policies in the United States. *National Transportation Policy* (Doyle Report). 87th Congress, 1st Session, 1960. Washington, D. C.: U. S. Government Printing Office, 1961.
12. Wilson, G. *Essays on Some Unsettled Questions in the Economics of Transportation.* Bloomington, Ind.: Foundation for Economic and Business Studies, Indiana University, 1962.